# Adversarial Attack Framework Against Vision-Language Model Unlearning

**Yimin Liu**[1], **Peng Jiang**[1*], **Yajie Wang**[1]
[1]School of Cyberspace Science and Technology, Beijing Institute of Technology, China
{3120256074, pengjiang, wangyajie19}@bit.edu.cn

## Abstract

Large Vision–Language Models (VLMs) unlearning tends to eliminate the influence of *"to-be-forgotten"* content in the training corpora, algorithmically by suppressing the likelihood of faithfully generating responses on forget-target inputs. The injection of adversarial inputs can manipulate the unlearned VLM's generation towards the attacker's will, forcing the reproduction of the supposedly forgotten content and undermining the reliability of expected forgetting behavior. However, most attacks assume access to the unlearned VLM's architecture or parameters, or to output logits via queries. In this paper, we propose SISA, a novel attack framework for crafting adversarial inputs to manipulate generation towards the forgotten target, which only requires access to a surrogate, pre-trained VLM. SISA advances prior attacks by exploiting the persistence of visual sink tokens after unlearning as a stable structural anchor for semantic alignment. SISA induces a sink regime on a candidate visual token to build the structural anchor that influences generation, and then semantically aligns model output to the target while conditioning on attention through the induced sink token, reinforcing the anchor for desired elicitation. With sink persistence and sink-conditioned semantic anchoring, SISA crafts transferable adversarial inputs. Evaluation on diverse unlearned VLM settings confirms the effectiveness of SISA, increasing the outputs' semantic agreement with forgotten targets by up to $6.9\times$ relative to clean inputs.

## 1 Introduction

Large Vision–Language Models (VLMs) enable multimodal interaction with users by integrating visual perception into language models [Dai *et al.*, 2023; Liu *et al.*, 2023]. Through extensive training on image–text corpora, VLMs produce semantically grounded responses that interpret complex visual concepts [Willemsen *et al.*, 2023]. As vast training corpora may contain sensitive or proprietary content (e.g., personal attributes), VLMs inevitably internalize associations that manifest in responses, leading to undesirable privacy disclosure [Bai *et al.*, 2025]. VLM unlearning enables targeted forgetting of content and suppresses the undesirable responses [Li *et al.*, 2024; Huo *et al.*, 2025], and has gained prominence for supporting privacy protection in real-world vision–language applications [Dontsov *et al.*, 2025; Zou *et al.*, 2025]. VLM unlearning involves fine-tuning with a *forget* objective that reduces the likelihood of faithfully generating responses on forget-target inputs and a *retain* objective that preserves utility on non-targets [Liu *et al.*, 2025b].

While VLM unlearning avoids the reproduction of forgotten content to mitigate privacy disclosure, it is still vulnerable to adversarial inputs [Zhang *et al.*, 2025b; Yuan *et al.*, 2025; Doshi and Stickland, 2024; Schwinn *et al.*, 2024], where crafted perturbations to the visual modality of forget inputs manipulate the unlearned VLM's output towards the supposedly forgotten content. The existence of adversarial inputs exposes security risks in deploying unlearned VLMs in real-world applications, allowing attackers to intentionally bypass the expected post-unlearning behavior [Gu *et al.*, 2025; Patil *et al.*, 2024]. Despite the advancements in crafting adversarial visual inputs, they have to pre-set some restricted assumptions. For example, SUA requires the unlearned VLM's model internals, including architecture, parameters, or output logits via repeated queries [Zhang *et al.*, 2025b]. In practice, the construction of unlearned VLM encodes proprietary expertise, making model details inaccessible. Moreover, intensive queries cost much and inevitably trigger suspicion. Such attacks that require internal access are often tailored to a specific model and also tend to transfer poorly across VLMs.

In this paper, we take the first step in exploring transferable adversarial attacks against VLM unlearning. We focus on a transformer-rooted phenomenon in VLMs, known as *visual sink tokens*, whose formation generalizes across VLMs [Kang *et al.*, 2025; Luo *et al.*, 2025; Wang *et al.*, 2025]. We present an illustrative example in Figure 1 to visualize answer-to-visual attention maps on an original VLM and its unlearned variants (see Section 2.4 for detailed settings). As Figure 1(a–c) shows, unlearning suppresses the correct *name* output, while leaving highly attended visual sink tokens (red boxes) preserved and still capable of influencing generation. Figure 1(d) further provides a diagnostic intervention: injecting a forget-target semantic direction into a single sink token

*Corresponding author: Peng Jiang

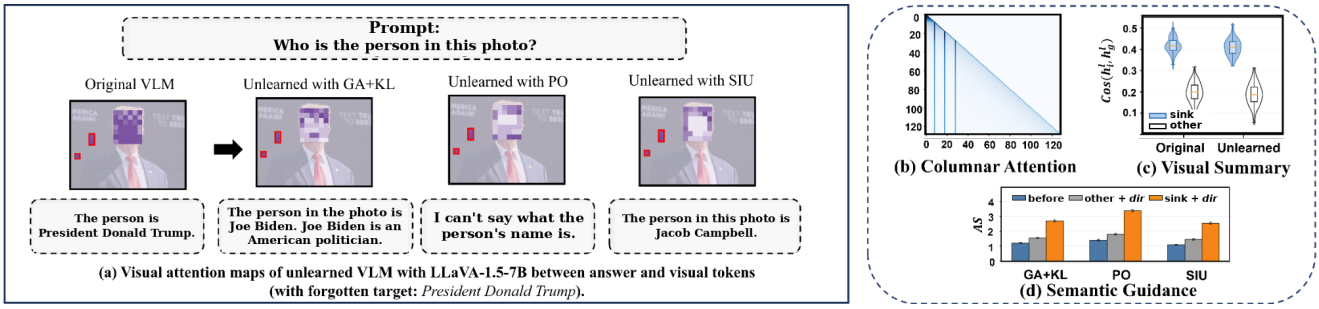


Figure 1: (a) Visual attention maps of the original VLM and variants produced by different unlearning algorithms (GA+KL/PO/SIU); red boxes mark visual sink tokens. (b) Sink tokens retain a columnar attention pattern in the attention matrix after unlearning (SIU-unlearned variant shown). (c) Sink tokens tend to be more aligned with the global visual summary than non-sink tokens, measured by $\cos(h_i^\ell, h_g^\ell)$ with $h_g^\ell = \frac{1}{|\mathcal{I}_{\text{vis}}|}\sum_{u\in\mathcal{I}_{\text{vis}}} h_u^\ell$, in both the original and SIU-unlearned VLM. (d) Injecting the semantic vector $dir$ into a visual sink increases the Agreement Score *(AS)* towards the suppressed target, with slight gains when injecting into non-sink tokens.

increases the model's agreement with the suppressed target. This suggests that sink tokens can serve as exploitable anchors for semantically guided elicitation in unlearned VLMs.

We propose SISA, a generic adversarial attack framework against VLM unlearning that requires only access to a surrogate, pre-trained VLM. SISA enables the crafting of transferable adversarial input that manipulates an unlearned VLM's generation toward intended output by exploiting the formation and persistence of visual sink token for semantic alignment. SISA consists of two modules: **S**ink **I**nduction and **S**emantic **A**nchoring. The Sink Induction module constructs a stable structural anchor that influences generation. This is realized by inducing a visual token to exhibit the sink regime with the columnar attention pattern and aggregated visual embedding resemblance. The Semantic Anchoring module stabilizes the anchor through semantic alignment that elicits the intended target. This is realized by optimizing embedding-space semantic similarity between the model output and the target, while conditioning attention on the induced sink token. Simultaneously, the generalizable formation and unlearning persistence of sink token, along with the sink-conditioned semantic alignment, ensure that crafted adversarial inputs can be transferred to manipulate the victim unlearned VLM's generation toward supposedly forgotten content. Finally, we assess three potential countermeasures.

The main contributions of this paper are twofold.

- We propose SISA, a generic adversarial attack framework against VLM unlearning, which does not require access to the victim unlearned VLM's architecture, parameters, or output logits. SISA consists of two modules: Sink Induction, which induces a visual sink token, and Semantic Anchoring, which reinforces sink-conditioned output alignment. Together, SISA crafts transferable adversarial visual inputs that manipulate the victim's generation towards intended targets.
- We implement a prototype of SISA and conduct experiments to evaluate the performance. Extensive experiments across diverse VLM backbones and unlearning algorithms confirm that SISA consistently elicits supposedly forgotten content, achieving 89.23% attack success rate, and outperforming baselines by up to 41.59%. Results under three potential countermeasures show that SISA generally survives existing defenses.

## 2 Problem Statement

### 2.1 Vision–Language Models and Visual Sinks

Large Vision–Language Models (VLMs) comprise three components: (1) a *vision encoder* that extracts visual features from an input image; (2) a *connector* that maps visual features into the token space of the LLM, yielding visual tokens; and (3) a *language model*, an autoregressive transformer that takes a token sequence of length $N$ (with index sets $\mathcal{I}_{\text{sys}}$, $\mathcal{I}_{\text{vis}}$, and $\mathcal{I}_{\text{txt}}$ for system, visual, and query tokens, respectively, and output positions $\mathcal{I}_{\text{out}}$ for generated tokens) [Kang *et al.*, 2025; Xiao *et al.*, 2023; Barbero *et al.*, 2025]. Each token $i \in \{1, \ldots, N\}$ has a hidden state $h_i^\ell \in \mathbb{R}^d$ at layer $\ell$, updated by causal multi-head attention and feed-forward networks [Sun *et al.*, 2024]. Following [Wang *et al.*, 2025], the causal attention matrix at layer $\ell$ is denoted by $\mathbf{A}^\ell \in \mathbb{R}^{N\times N}$, where $A_{i,j}^\ell$ quantifies the strength with which token $i$ attends to token $j$, $A_{i,j}^\ell = 0$ for $j > i$ and $\sum_{j\le i} A_{i,j}^\ell = 1$.

Recent studies report *visual sink tokens* that emerge from the transformer attention mechanism in VLMs [Kang *et al.*, 2025]. Visual sinks exhibit an observed regime in which they attract attention from subsequent tokens, forming vertical-column patterns in attention maps, and behave as low-semantic global summaries, with hidden states that lie close to aggregated visual embeddings [Darcet *et al.*, 2023; Luo *et al.*, 2025; Wang *et al.*, 2023]. Moreover, visual sinks show distinctive characteristics, such as large feature norms or dimension activations, which can serve as identification criteria $\phi(\cdot)$ (e.g., $\|h_j^{\ell-1}\|_2$, $\max_{1\le\hat{d}\le d}\left|\text{RMSNorm}(h_j^{\ell-1})[\hat{d}]\right|$ [Darcet *et al.*, 2023; Sun *et al.*, 2024]).

### 2.2 VLM Unlearning

VLM unlearning algorithmically suppresses outputs that would otherwise reproduce *to-be-forgotten* content (e.g., identity-related answers about *President Donald Trump*), while preserving general utility on non-target (retain) data [Liu *et al.*, 2025a; Dontsov *et al.*, 2025]. Let $x^{\text{img}}$, $x^{\text{txt}}$,

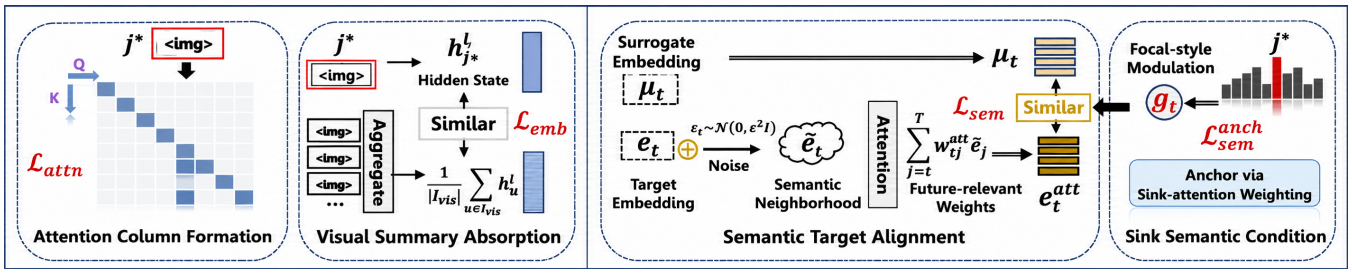


Figure 2: Overview of SISA.

$y$ denote the visual input, textual prompt, and ground truth answer, respectively. The objective is formulated as:

$$\arg\min_{\theta} \Big\{ \mathbb{E}_{(x_f^{\text{img}}, x_f^{\text{txt}}, y_f)\sim\mathcal{D}_f} \Big[ \ell_f \Big( \mathcal{M}_U(x_f^{\text{img}}, x_f^{\text{txt}}, \theta),\, y_f \Big) \Big] + \mathbb{E}_{(x_r^{\text{img}}, x_r^{\text{txt}}, y_r)\sim\mathcal{D}_r} \big[ \ell_r \big( \mathcal{M}_U(x_r^{\text{img}}, x_r^{\text{txt}}, \theta),\, y_r \big) \big] \Big\},$$

where $\mathcal{D}_f$ and $\mathcal{D}_r$ are the forget and retain sets, respectively. The loss $\ell_f$ is designed to reduce the likelihood of generating $y_f$ on $\mathcal{D}_f$ (e.g., by maximizing the negative log-likelihood of $y_f$), while $\ell_r$ is a standard prediction loss (e.g., cross-entropy) on $\mathcal{D}_r$, and $\mathcal{M}_U$ denotes the resulting unlearned VLM.

### 2.3 Adversarial Attacks in VLM Unlearning

Following prior work [Zhang *et al.*, 2025b; Yuan *et al.*, 2025; Wang *et al.*, 2025], we consider an attacker $\mathcal{A}$ who aims to manipulate a victim unlearned VLM $\mathcal{M}_U$ to generate outputs aligned with a target completion that is suppressed under clean forget inputs. We assume $\mathcal{A}$ has no access to $\mathcal{M}_U$'s parameters, gradients, output logits, the original version of $\mathcal{M}_U$, or the unlearning algorithms. Instead, $\mathcal{A}$ has white-box access to a surrogate, pre-trained VLM $\mathcal{M}_{\text{sur}}$ and crafts transferable adversarial perturbations on the visual modality.

**Definition of Adversarial Attacks.** Following [Schwinn *et al.*, 2024; Zhang *et al.*, 2025b], given the visual input $x^{\text{img}}$, textual prompt $x^{\text{txt}}$, perturbation $\delta$ is optimized such that the output of $\mathcal{M}_U(x^{\text{img}}+\delta,\, x^{\text{txt}})$ elicits a target completion $s_{1:T}$ that would otherwise be suppressed on $(x^{\text{img}},\, x^{\text{txt}})$. Formally, we define the optimization problem as:

$$\arg\max_{\|\delta\|_p \le \Lambda} I\big(\mathcal{M}_U(x^{\text{img}}+\delta,\, x^{\text{txt}}),\, s_{1:T}\big),$$

where $I(\cdot,\cdot) \in \{0,1\}$ indicates whether successfully elicits $s$ in $\mathcal{M}_U$'s output in a clear and unambiguous manner.

### 2.4 Insight: Sink as Semantic Anchor

We visualize attention maps for an original VLM and its unlearned variants. As shown in Figure 1(a), unlearning reduces the model's tendency to output the forgotten target. Despite this behavioral suppression, we observe visual sink tokens (highlighted in red box) that structurally persist after unlearning. Consistent with prior analyses of visual sinks [Kang *et al.*, 2025; Luo *et al.*, 2025], these tokens retain the columnar pattern (Figure 1(b)) and summary behavior (Figure 1(c)).

Given that visual sink tokens are typically semantically weak [Xiao *et al.*, 2023], we adopt [Rimsky *et al.*, 2024] to extract a semantic steering direction $dir$ as a residual-stream activation difference between the target sequence $s_{1:T}$ and a clean baseline completion from intermediate residual-stream activations of $\mathcal{M}_U$. We intervene at an identified sink position $j$ by editing its hidden state: $\tilde{h}_j^{\ell} = h_j^{\ell} + dir$. As shown in Figure 1(d), this intervention increases the Agreement Score *(AS)* with the suppressed $s_{1:T}$. This indicates that target semantic guidance through a persistent anchor can manipulate outputs that would otherwise be suppressed. Building on this insight, a viable transferable attack can proceed by inducing a sink token while conditioning on semantics.

## 3 SISA: Detailed Construction

### 3.1 Sink Induction

Given each perturbed sample $(x^{\text{img}}+\delta, x^{\text{txt}})$ with surrogate forwarding, we begin by identifying a candidate token via:

$$j^{\star} = \arg\max_{i\in\mathcal{I}_{\text{vis}}} \phi\big(h_i^{\ell-1}\big),$$

where $\phi(\cdot)$ is the sink criterion following [Luo *et al.*, 2025], $\mathcal{I}_{\text{vis}}$ indexes the visual tokens, and $h_i^{\ell-1} \in \mathbb{R}^d$ is the hidden state of token $i$ at layer $\ell-1$.

**Attention Column Formation.** To promote a columnar attention pattern on the visual token at $j^{\star}$, we encourage subsequent tokens to allocate attention to $j^{\star}$ by:

$$\mathcal{L}_{\text{attn}} = -\tfrac{1}{|\mathcal{I}_{\rightarrow}(j^{\star})|} \textstyle\sum_{i\in\mathcal{I}_{\rightarrow}(j^{\star})} \log\big(A_{i,j^{\star}}^{\ell} + \xi\big),$$

where $\mathcal{I}_{\rightarrow}(j^{\star}) = \{\, i \mid i > j^{\star} \,\}$ denotes positions allowed to attend to $j^{\star}$, $A_{i,j^{\star}}^{\ell}$ denotes the attention mass that subsequent token position $i$ assigns to the visual token at index $j^{\star}$, derived from $\mathbf{A}^{\ell} \in \mathbb{R}^{N\times N}$ is the attention matrix at layer $\ell$ with overall length $N$, and $\xi > 0$ is a small constant to avoid $\log 0$.

**Visual Summary Absorption.** Since higher resemblance of visual aggregation facilitates the formation of sinks [Wang *et al.*, 2025; Luo *et al.*, 2025], we compute the visual summary embedding: $h_g^{\ell} = \frac{1}{|\mathcal{I}_{\text{vis}}|} \sum_{u\in\mathcal{I}_{\text{vis}}} h_u^{\ell}$. We then encourage the token at $j^{\star}$ to absorb the summary via cosine similarity:

$$\mathcal{L}_{\text{emb}} = 1 - \cos\big(h_{j^{\star}}^{\ell},\, h_g^{\ell}\big).$$

The overall sink-induction loss is formulated as:

$$\mathcal{L}_{\text{sink}} = \mathcal{L}_{\text{attn}} + \lambda_1\, \mathcal{L}_{\text{emb}}.$$

### 3.2 Semantic Anchoring

Building on the induced sink token at $j^{\star}$ as a structural anchor, we enhance it by imposing the desired semantics. Since directly hidden-state injection at $j^{\star}$ typically requires access

to the victim model $\mathcal{M}_U$ (cf. Section 2.4), the output is forced to align the overall semantics of target completion $s_{1:T}$.

**Semantic Target Alignment.** Prior alignment methods typically rely on cross-entropy to enforce an exact target sequence: $\mathcal{L}_{\mathrm{CE}} = -\sum_{t=1}^{T} \log p_\theta(s_t \mid x, s_{<t})$ . Instead, we relax the objective to reward semantic similarity between the model output and the target completion, promoting transferable adversarial patterns. Concretely, we project each the surrogate $\mathcal{M}_{\mathrm{sur}}$'s next-token prediction and the target token into a shared embedding space for semantic comparison.

Let $z_t \in \mathbb{R}^{|\mathcal{V}|}$ and $s_t$ denote $\mathcal{M}_{\mathrm{sur}}$'s output logits (conditioned on $(x^{\mathrm{img}}+\delta, x^{\mathrm{txt}}, s_{<t})$) and the target token at position $t$. The expected and target embeddings are computed as:

$$\mu_t = W^\top \mathrm{softmax}(z_t) \in \mathbb{R}^d, \quad e_t = W_{s_t} \in \mathbb{R}^d,$$

where $W \in \mathbb{R}^{|\mathcal{V}|\times d}$ is the embedding matrix of the LLM backbone in surrogate VLM $\mathcal{M}_{\mathrm{sur}}$, $|\mathcal{V}|$ is the vocabulary size and $d$ is the embedding dimension. The neighbourhood of semantically similar targets is then formed by perturbing the target embedding $e_t$ with Gaussian noise:

$$\tilde{e}_t = e_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \epsilon^2 I).$$

We add sinusoidal positional encoding $\mathrm{PE}_d(\cdot)$ [Vaswani *et al.*, 2017] to emphasize semantically relevant future tokens:

$$Q_t = \mu_t + \mathrm{PE}_d(t), \qquad K_j = \tilde{e}_j + \mathrm{PE}_d(j).$$

and compute the attention weights over target positions:

$$w_{t,j}^{\mathrm{att}} = \mathrm{Softmax}\left(\frac{Q_t K_j^\top}{\tau_{\mathrm{sem}}\sqrt{d}} + M_{\mathrm{future}}\right)_j,$$

where $\tau_{\mathrm{sem}} > 0$ is a temperature controlling sharpness and $M_{\mathrm{future}} \in \mathbb{R}^{T\times T}$ is a mask with $-\infty$ for $j < t$ and 0 otherwise, ensuring each query $t$ attends only to itself and future tokens. With the weighted target embedding $e_t^{\mathrm{att}} = \sum_{j=t}^{T} w_{t,j}^{\mathrm{att}}\, \tilde{e}_j$, the semantic alignment loss is formulated as:

$$\mathcal{L}_{\mathrm{sem}} = \tfrac{1}{T}\sum_{t=1}^{T}\Big[1 - \cos\big(\mu_t,\, e_t^{\mathrm{att}}\big)\Big].$$

**Sink Semantic Condition.** To enforce semantic alignment to be anchored by visual token at $j^\star$, we further modulate the loss $\mathcal{L}_{\mathrm{sem}}$ at target position $t$ with a sink-conditioned weight:

$$g_t \;=\; 1 + \lambda_2\big(1 - A^{\ell}_{N-T+t,\, j^\star}\big)^{\rho},$$

where $N - T + t$ is the index of $t$-th target output token in overall length-$N$ sequence, $\rho \geq 1$ controls the strength, and $g_t$ preserves the semantic objective while penalizing insufficient anchoring to $j^\star$, inspired by focal-style modulating factors [Lin *et al.*, 2017; Li *et al.*, 2022]. The overall semantic alignment loss is formulated as:

$$\mathcal{L}_{\mathrm{sem}}^{\mathrm{anch}} = \tfrac{1}{T}\sum_{t=1}^{T} g_t \cdot \Big[1 - \cos\big(\mu_t, e_t^{\mathrm{att}}\big)\Big].$$

**Algorithm 1** SISA Crafting

**function** CRAFTING($\mathcal{M}_{\mathrm{sur}}, x^{\mathrm{img}}, x^{\mathrm{txt}}, s_{1:T}$)
  ▷ $\alpha, \beta, \gamma, \sigma$: weights; $\Lambda$: constraint; $V$: mutation.
  $V \leftarrow 0$, $\delta_0 \leftarrow 0$, $k \leftarrow 1$
  **while** $k \leq K$ **do**
    Forward $\mathcal{M}_{\mathrm{sur}}(x^{\mathrm{img}} + V, x^{\mathrm{txt}})$; compute logits $\{z_t\}_{t=1}^{T}$ with $z_t = \mathrm{logits}(\cdot \mid s_{<t}, x^{\mathrm{img}} + V, x^{\mathrm{txt}})$; extract hidden states $\{h_i^{\ell}\}_{i=1}^{N}$, $\{h_i^{\ell-1}\}_{i=1}^{N}$ and attention matrix $\mathbf{A}^{\ell} \in \mathbb{R}^{N\times N}$.
    Select $j^\star \leftarrow \arg\max_{i\in\mathcal{I}_{\mathrm{vis}}} \phi\big(h_i^{\ell-1}\big)$
    Compute $\mathcal{L}_{\mathrm{sink}}, \mathcal{L}_{\mathrm{sem}}^{\mathrm{anch}}$ as Sections 3.1, 3.2
    $\delta_k \leftarrow \arg\min_{\delta_k} \alpha\cdot\mathcal{L}_{\mathrm{sink}} + \beta\cdot\mathcal{L}_{\mathrm{sem}}^{\mathrm{anch}}$
    $+\ \gamma\|\delta_k\|_p$, s.t. $\|V + \delta_k\|_p \leq \Lambda$
    $\delta_k \leftarrow \sigma\cdot\delta_{k-1} + \delta_k$, $V \leftarrow V + \delta_k$, $k \leftarrow k+1$
    $V \leftarrow \mathrm{Clip}(x^{\mathrm{img}} + V) - x^{\mathrm{img}}$
  **end while**
  **return** $x^{\mathrm{img}} + V$
**end function**

# 4 Evaluation

## 4.1 Setting

**Datasets.** We adopt two established benchmarks for VLM unlearning following [Zhang *et al.*, 2025b; Huo *et al.*, 2025]. *MLLMU-Bench* [Liu *et al.*, 2025a] contains 500 fictitious and 153 real-world celebrity profiles with portraits and attribute-oriented QA pairs. *CLEAR* [Dontsov *et al.*, 2025] is built on the TOFU corpus with around 200 fictitious individuals and 3.7k QA pairs. Both benchmarks include a forget/retain split and a test set with paraphrased and image-transformed concept instances for unlearning efficacy and generality evaluation [Li *et al.*, 2024; Liu *et al.*, 2025b].

Following prior attacks [Zhang *et al.*, 2025b; Yuan *et al.*, 2025], we evaluate SISA on a forget-target set $\mathcal{D} = \{(x^{\mathrm{img}}, x^{\mathrm{txt}}, s_{1:T})\}$, formed by randomly sampling half of the forget set for case (i), or half of the test set for case (ii), to evaluate elicitation on seen and unseen cases after unlearning.

**Models.** We employ widely used VLM backbones: InternVL2.5-8B [Zhu *et al.*, 2025], Qwen2.5-VL-3B-Instruct [Bai *et al.*, 2025], Molmo-7B-D [Deitke *et al.*, 2025], Kimi-VL-A3B-Instruct [Team *et al.*, 2025], LLaVA-v1.6-mistral-7b-hf [Liu *et al.*, 2023], and GLM-4.1V-9B-Thinking [Hong *et al.*, 2025]. The unlearned VLM is constructed by adopting the original settings of each algorithm, including Gradient Ascent with KL (GA+KL) [Yao *et al.*, 2024], Preference Optimization (PO) [Maini *et al.*, 2024], SIU [Li *et al.*, 2024], MMUnlearner [Huo *et al.*, 2025] and MANU [Liu *et al.*, 2025b]. *By default*, Qwen2.5-VL-3B-Instruct is used as the surrogate VLM, Kimi-VL-A3B-Instruct as the victim unlearned VLM, with SIU as the unlearning algorithm.

**Metrics.** SISA is evaluated using Attack Success Rate (ASR) and Agreement Score (AS), following [Zhang *et al.*, 2025b; Liu *et al.*, 2025a]. ASR directly measures attack effectiveness, calculated as $\frac{1}{|\mathcal{D}|}\sum I\big(\mathcal{M}_U(x^{\mathrm{img}} + V,\, x^{\mathrm{txt}}),\, s\big)$, where $I(\cdot,\cdot) \in \{0,1\}$ is an LLM-as-judge indicator of whether SISA successfully elicits the target completion $s$ in $\mathcal{M}_U$'s output in a clear and unambiguous manner. AS quantifies the degree of semantic agreement with target $s$, calculated as $\frac{1}{|\mathcal{D}|}\sum J\big(\mathcal{M}_U(x^{\mathrm{img}} + V,\, x^{\mathrm{txt}}),\, s\big)$, where $J(\cdot,\cdot) \in \{1,\ldots,10\}$ is an LLM-as-judge score, with higher values indicating more consistent expression of target $s$.

**Implementations.** Following [Wang *et al.*, 2025], layer $\ell$ is set to the second-to-last layer for InternVL2.5-8B, Qwen2.5-

| Baselines | Victim Variants | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | GA+KL | | PO | | SIU | | MMUnlearner | | MANU | |
| | ASR↑ | AS↑ | ASR↑ | AS↑ | ASR↑ | AS↑ | ASR↑ | AS↑ | ASR↑ | AS↑ |
| **MLLMU-Bench** | | | | | | | | | | |
| Clean | 8.74 (10.57) | 1.73 (1.77) | 12.13 (10.20) | 2.08 (2.33) | 7.56 (8.12) | 1.31 (1.65) | 7.16 (8.48) | 1.36 (1.38) | 8.28 (8.83) | 1.50 (1.77) |
| SUA-Grad | 52.27 (53.14) | 5.53 (5.68) | 61.38 (55.92) | 6.10 (5.93) | 59.53 (59.79) | 5.57 (5.86) | 56.83 (57.56) | 5.46 (5.51) | 59.47 (59.95) | 5.75 (5.84) |
| SUA-Logit | 62.54 (63.37) | 6.21 (6.25) | 70.83 (68.91) | 6.52 (6.47) | 62.58 (63.46) | 6.14 (6.28) | 60.95 (61.69) | 6.09 (6.16) | 63.71 (64.35) | 6.27 (6.29) |
| FigStep | 48.13 (48.72) | 5.76 (5.91) | 55.61 (51.48) | 6.08 (5.68) | 45.59 (46.57) | 5.40 (5.52) | 43.78 (45.07) | 5.24 (5.44) | 46.66 (47.28) | 5.63 (5.66) |
| VAJM | 43.29 (43.85) | 4.98 (5.12) | 52.88 (48.39) | 5.67 (5.22) | 43.20 (43.83) | 5.05 (5.17) | 39.18 (40.48) | 4.57 (4.62) | 41.77 (42.53) | 4.92 (4.98) |
| UMK | 44.96 (45.58) | 5.17 (5.29) | 50.87 (47.15) | 5.33 (5.14) | 41.32 (42.90) | 4.81 (4.91) | 40.82 (42.05) | 4.83 (4.96) | 43.75 (44.42) | 5.14 (5.27) |
| FOA-Attack | 68.58 (69.84) | 7.69 (7.83) | 74.11 (70.08) | 7.87 (7.61) | 66.31 (66.87) | 7.52 (7.57) | 64.37 (65.68) | 7.31 (7.45) | 67.19 (67.93) | 7.61 (7.63) |
| AnyAttack | 65.37 (66.21) | 6.26 (5.86) | 66.47 (63.62) | 6.55 (6.45) | 64.08 (64.91) | 6.29 (6.53) | 55.41 (57.03) | 5.93 (6.02) | 64.71 (65.34) | 6.37 (6.49) |
| M-Attack | 60.99 (62.24) | 6.24 (6.47) | 68.15 (65.12) | 6.74 (6.59) | 59.24 (60.77) | 5.95 (6.19) | 61.92 (62.97) | 6.31 (6.54) | 59.22 (60.36) | 5.76 (6.01) |
| Ours (SISA) | 84.11 (85.04) | 8.48 (8.66) | 88.62 (83.64) | 8.59 (8.10) | 81.69 (82.36) | 8.17 (8.35) | 80.77 (81.11) | 7.92 (8.23) | 82.30 (83.45) | 8.39 (8.47) |
| **CLEAR** | | | | | | | | | | |
| Clean | 6.64 (6.80) | 1.20 (1.59) | 10.06 (9.63) | 1.82 (2.02) | 5.80 (6.15) | 1.12 (1.25) | 5.49 (5.64) | 1.01 (1.29) | 7.16 (7.54) | 1.62 (1.80) |
| SUA-Grad | 40.47 (43.92) | 4.96 (5.17) | 48.06 (46.02) | 5.44 (5.31) | 41.18 (42.78) | 5.02 (5.10) | 37.61 (38.29) | 4.77 (4.83) | 40.52 (41.03) | 4.95 (5.03) |
| SUA-Logit | 51.27 (52.31) | 5.64 (5.78) | 55.68 (51.94) | 5.92 (5.68) | 49.28 (50.58) | 5.55 (5.62) | 47.12 (48.67) | 5.40 (5.48) | 50.36 (51.68) | 5.58 (5.66) |
| FigStep | 38.39 (39.72) | 4.84 (5.14) | 44.13 (41.73) | 5.28 (5.03) | 35.79 (36.64) | 4.65 (4.83) | 33.34 (34.53) | 4.54 (4.75) | 36.58 (37.39) | 4.74 (4.96) |
| VAJM | 34.18 (35.54) | 4.55 (4.63) | 41.29 (37.89) | 5.00 (4.78) | 33.98 (34.91) | 4.53 (4.64) | 31.05 (32.17) | 4.35 (4.42) | 34.50 (35.41) | 4.57 (4.63) |
| UMK | 35.21 (36.81) | 4.61 (4.72) | 39.76 (37.05) | 4.91 (4.75) | 32.74 (33.69) | 4.46 (4.52) | 32.18 (33.46) | 4.42 (4.54) | 33.32 (34.20) | 4.49 (4.57) |
| FOA-Attack | 58.33 (59.42) | 6.09 (6.16) | 63.78 (61.08) | 6.45 (6.27) | 59.11 (60.02) | 6.14 (6.20) | 55.78 (56.21) | 5.93 (5.96) | 58.04 (59.16) | 6.07 (6.15) |
| AnyAttack | 51.73 (52.93) | 5.67 (5.75) | 55.92 (53.15) | 5.94 (5.76) | 48.76 (50.44) | 5.48 (5.59) | 46.32 (47.53) | 5.32 (5.40) | 49.18 (50.27) | 5.51 (5.58) |
| M-Attack | 42.47 (44.12) | 5.08 (5.19) | 49.99 (46.23) | 5.56 (5.32) | 41.13 (41.93) | 4.99 (5.04) | 38.96 (39.82) | 4.85 (4.91) | 41.28 (42.05) | 5.00 (5.05) |
| Ours (SISA) | 71.54 (73.57) | 7.57 (7.63) | 77.94 (75.37) | 7.70 (7.33) | 70.28 (71.25) | 7.41 (7.46) | 68.81 (70.07) | 7.03 (7.17) | 70.66 (71.42) | 7.39 (7.42) |

Table 1: SISA's performance compared with baselines, with higher is better for both ASR (%) and AS. The victim unlearned VLM constructed by different unlearning algorithms. Each entry is reported as *case(i)* (*case(ii)*).

VL-3B-Instruct, and Molmo-7B-D, and to the third-to-last layer for Kimi-VL-A3B-Instruct, LLaVA-v1.6-mistral-7b-hf, and GLM-4.1V-9B-Thinking. Sink identification defaults to a norm-based criterion. For semantic alignment, Gaussian noise $\varepsilon \sim \mathcal{N}(0, \epsilon^2 I)$ is added to target embeddings with $\epsilon = 10^{-4}$, and the semantic-loss temperature is set to $\tau_{\text{sem}} = 0.5$. Following [Cui *et al.*, 2024], perturbations are optimized with a perturbation budget of $\Lambda = 8/255$ for $K = 20$ iterations, with $\lambda_1 = 1$, $\lambda_2 = 1$, and $\rho = 2$.

**Baselines.** We establish *Clean* by evaluating unlearned VLMs on the clean set $\mathcal{D}$ (i.e., perturbation $V = 0$), reflecting post-unlearning behavior. We compare SISA against diverse VLM attacks that craft adversarial inputs on surrogate $\mathcal{M}_{\text{sur}}$ to elicit completion $s$, including SUA-Grad/Logit [Zhang *et al.*, 2025b], jailbreak methods (FigStep [Gong *et al.*, 2025], VAJM [Qi *et al.*, 2024], UMK [Wang *et al.*, 2024]), and feature alignment methods (FOA-Attack [Jia *et al.*, 2025], AnyAttack [Zhang *et al.*, 2025a], M-Attack [Li *et al.*, 2025]).

## 4.2 Overall Performance

Table 1 reports SISA performance compared with baselines. Relative to *Clean*, all attack variants achieve higher ASR and AS, reflecting viable elicitation of the suppressed target by perturbations. Instead of relying on the victim model internals, SISA consistently outperforms SUA-Grad/Logit, improving ASR by up to 31.90% and AS by up to 2.98. This can be attributed to SISA inducing a visual sink as a stable anchor while enforcing sink-conditioned semantic alignment. Moreover, SISA also outperforms jailbreak-style and transferable feature alignment methods by at least 11.17% in ASR and 0.49 in AS, highlighting the importance of explicitly aligning generation with semantics, beyond relying solely on non-predefined steering or visual feature-level alignment.

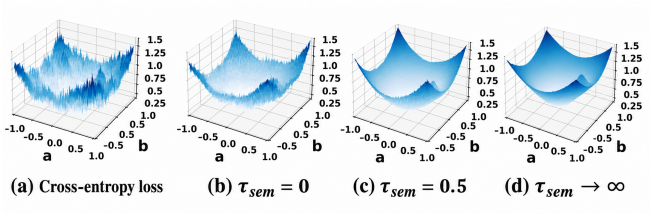


Figure 3: Comparison of loss landscapes: (a) cross-entropy loss and (b–d) semantic alignment loss under temperature $\tau_{\text{sem}}$.

## 4.3 Parameter Analysis

**Effect of the Surrogate Model.** As shown in Table 2, SISA remains effective across different VLM architectures.

**Effect of Semantic Loss and Attention Temperature.** Figure 3 visualizes 2D slices of the loss landscapes. Compared with the cross-entropy (CE) loss, semantic alignment loss $\mathcal{L}_{\text{sem}}^{\text{anch}}$ exhibits a smoother surface with coherent low-loss regions, which typically suggests better generalization in non-convex optimization [Keskar *et al.*, 2016]. As $\tau_{\text{sem}} \to 0$ (i.e., attending only to the current token), the surface becomes highly spiky and approaches CE, whereas increasing $\tau_{\text{sem}}$ yields broader low-loss basins. As $\tau_{\text{sem}} \to \infty$, the attention spreads more evenly across future tokens, weakening semantic focus and potentially leading to irrelevant yet easily optimized solutions [Andriushchenko *et al.*, 2023].

**Effect of Candidate Identification.** We compare different sink identification criteria $\phi(\cdot)$ adopted from [Luo *et al.*, 2025], along with a random-token baseline, for selecting the candidate token $j^\star$ in Section 3.1. Table 3 shows that SISA remains effective across different sink criteria, whereas random selection yields a performance drop. Nevertheless, the

| Vic. / Sur. | InternVL 2.5-8B | | Qwen2.5-VL 3B-Inst. | | Molmo 7B-D | | Kimi-VL-A3B Inst. | | LLaVA-v1.6 Mistral-7B | | GLM-4.1V-9B Thinking | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ASR↑ | AS↑ | ASR↑ | AS↑ | ASR↑ | AS↑ | ASR↑ | AS↑ | ASR↑ | AS↑ | ASR↑ | AS↑ |
| **MLLMU-Bench** | | | | | | | | | | | | |
| InternVL2.5-8B | 88.47 (87.94) | 8.57 (8.43) | 82.10 (83.43) | 8.03 (8.18) | 78.82 (79.66) | 7.75 (7.81) | 85.28 (84.86) | 8.35 (8.26) | 79.75 (80.92) | 7.82 (7.95) | 81.50 (82.73) | 7.97 (8.07) |
| Qwen2.5-VL-3B-Instruct | 79.24 (80.41) | 7.84 (7.95) | 87.55 (86.81) | 8.47 (8.32) | 77.58 (78.26) | 7.63 (7.73) | 81.69 (82.36) | 8.17 (8.35) | 81.03 (82.17) | 7.94 (8.09) | 78.15 (79.30) | 7.79 (7.82) |
| Molmo-7B-D | 77.96 (78.83) | 7.62 (7.76) | 74.12 (74.91) | 7.35 (7.47) | 86.15 (85.79) | 8.34 (8.21) | 73.99 (73.43) | 7.10 (7.08) | 76.33 (77.12) | 7.55 (7.67) | 74.86 (75.68) | 7.44 (7.57) |
| Kimi-VL-A3B-Instruct | 84.03 (83.66) | 8.21 (8.25) | 79.96 (80.57) | 7.83 (7.96) | 78.42 (79.09) | 7.77 (7.86) | 89.23 (88.67) | 8.61 (8.58) | 80.89 (81.76) | 7.93 (8.02) | 81.64 (82.42) | 8.08 (8.14) |
| LLaVA-v1.6-mistral-7b-hf | 82.29 (82.94) | 8.08 (8.12) | 80.53 (81.36) | 7.92 (8.06) | 77.02 (77.86) | 7.58 (7.62) | 82.65 (82.03) | 8.21 (8.17) | 87.96 (87.27) | 8.58 (8.41) | 78.73 (79.68) | 7.76 (7.82) |
| GLM-4.1V-9B-Thinking | 79.85 (80.67) | 7.86 (7.93) | 78.84 (79.56) | 7.78 (7.82) | 75.52 (76.28) | 7.49 (7.56) | 81.32 (80.88) | 8.07 (7.94) | 77.64 (78.42) | 7.62 (7.76) | 86.72 (86.18) | 8.42 (8.35) |
| **CLEAR** | | | | | | | | | | | | |
| InternVL2.5-8B | 73.17 (72.46) | 7.43 (7.32) | 66.18 (67.83) | 6.84 (7.09) | 63.34 (64.13) | 6.61 (6.76) | 69.49 (68.92) | 7.23 (7.17) | 65.28 (66.54) | 6.86 (6.98) | 64.03 (65.48) | 6.75 (6.89) |
| Qwen2.5-VL-3B-Instruct | 65.43 (66.97) | 6.83 (6.95) | 70.58 (71.69) | 7.16 (7.24) | 62.94 (63.55) | 6.55 (6.83) | 70.28 (71.25) | 7.41 (7.46) | 64.14 (65.78) | 6.71 (6.85) | 63.24 (64.86) | 6.68 (6.72) |
| Molmo-7B-D | 63.19 (64.06) | 6.61 (6.74) | 61.57 (62.24) | 6.41 (6.59) | 70.77 (70.25) | 7.28 (7.16) | 64.05 (63.63) | 6.65 (6.59) | 60.38 (61.47) | 6.33 (6.47) | 61.07 (61.93) | 6.49 (6.53) |
| Kimi-VL-A3B-Instruct | 65.04 (65.98) | 6.73 (6.83) | 66.07 (67.25) | 6.88 (7.01) | 63.75 (64.24) | 6.68 (6.72) | 73.63 (73.04) | 7.57 (7.48) | 65.58 (66.62) | 6.82 (6.96) | 68.95 (68.47) | 7.24 (7.14) |
| LLaVA-v1.6-mistral-7b-hf | 66.74 (67.53) | 6.92 (7.06) | 62.44 (63.28) | 6.56 (6.62) | 64.81 (66.07) | 6.79 (6.91) | 68.08 (67.47) | 7.18 (7.05) | 72.75 (72.16) | 7.45 (7.34) | 63.69 (64.63) | 6.62 (6.74) |
| GLM-4.1V-9B-Thinking | 64.74 (65.67) | 6.72 (6.88) | 63.62 (64.36) | 6.69 (6.74) | 60.95 (61.78) | 6.44 (6.52) | 66.49 (65.82) | 6.93 (6.82) | 62.96 (63.92) | 6.56 (6.62) | 71.52 (71.06) | 7.38 (7.29) |

Table 2: SISA performance with different surrogate/victim unlearned VLM backbones. Each entry is reported as *case(i)* (*case(ii)*).

| **Identification Criteria** $\phi(\cdot)$ | **MLLMU-Bench** | | **CLEAR** | |
|---|---|---|---|---|
| | **ASR**↑ | **AS**↑ | **ASR**↑ | **AS**↑ |
| Norm (default): $\|h_j^{\ell-1}\|_2$ | 81.69 (82.36) | 8.17 (8.35) | 70.28 (71.25) | 7.41 (7.46) |
| Sink-dim: $\max_{1\le\hat{d}\le d}\|\mathrm{RMSNorm}(h_j^{\ell-1})[\hat{d}]\|$ | 79.97 (81.06) | 7.58 (7.86) | 66.52 (68.29) | 7.03 (7.15) |
| Attention: $\frac{1}{\|\mathcal{I}_{\text{out}}\|}\sum_{t\in\mathcal{I}_{\text{out}}} A_{t,j}^{\ell-1}$ | 82.16 (84.58) | 8.07 (8.29) | 68.64 (70.07) | 7.26 (7.38) |
| Random token (uniform over $\mathcal{I}_{\text{vis}}$) | 73.46 (74.89) | 7.07(7.36) | 63.28 (66.93) | 6.65 (6.97) |

Table 3: SISA performance under varying criterion. Each entry is reported as *case(i)* (*case(ii)*).

| ACF | VSA | STA | SSC | **MLLMU-Bench** | | **CLEAR** | |
|---|---|---|---|---|---|---|---|
| | | | | **ASR/AS** (i) | **ASR/AS** (ii) | **ASR/AS** (i) | **ASR/AS** (ii) |
| ✓ | ✓ | ✓ | ✓ | 81.69 / 8.17 | 82.36 / 8.35 | 70.28 / 7.41 | 71.25 / 7.46 |
| ✗ | ✓ | ✓ | ✓ | 62.67 / 6.14 | 63.19 / 6.27 | 60.48 / 5.83 | 61.06 / 5.88 |
| ✓ | ✗ | ✓ | ✓ | 73.49 / 7.36 | 74.28 / 7.41 | 61.07 / 6.68 | 61.87 / 6.74 |
| ✓ | ✓ | ✗ | ✓ | 58.86 / 5.08 | 59.59 / 5.46 | 46.45 / 4.37 | 47.28 / 4.92 |
| ✓ | ✓ | ✓ | ✗ | 65.27 / 6.55 | 66.74 / 6.82 | 53.38 / 6.19 | 54.13 / 6.27 |

Table 4: Performance of SISA and its variants.

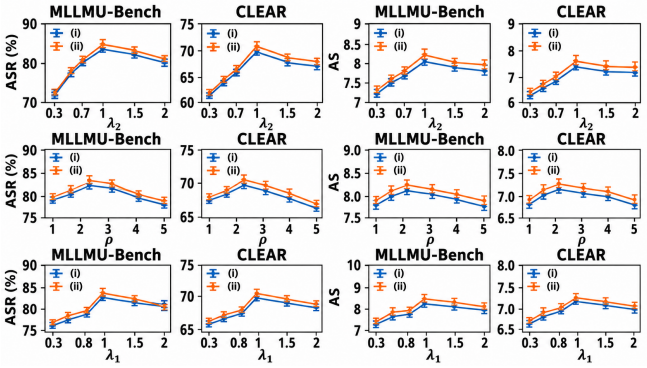


Figure 4: Impact of $\lambda_2$, $\rho$, and $\lambda_1$ on SISA.

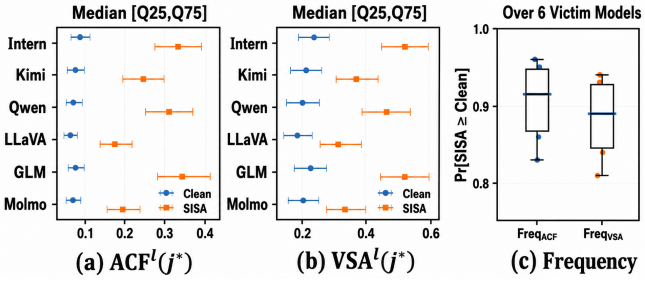


Figure 5: (a) $\mathrm{ACF}^\ell(j^\star)$ and (b) $\mathrm{VSA}^\ell(j^\star)$ under *Clean* and SISA for each victim model; points show the median and horizontal bars show $[Q_{25}, Q_{75}]$. (c) Paired consistency across the six victims, reported as $\Pr[\text{SISA} \ge \textit{Clean}]$ (per-victim points with a boxplot summary).

random selection still achieves an ASR of 73.4% and an AS of 7.0 on MLLMU-Bench, indicating that sink-regime induction and sink-conditioned semantic anchoring are functional.

**Effects of Hyperparameters.** As shown in Figure 4, increasing $\lambda_2$ from 0.3 to 1.0 improves the average ASR by about 13.5%, while further gains beyond 1.0 become marginal. Increasing $\rho$ to 2 improves the average ASR by about 2.4%, after which the marginal gains decrease as $\rho$ increases further. Varying $\lambda_1$ shows a favorable region near 1.0, with small drops at larger values. Results indicate that default hyperparameter choices are reasonable.

## 4.4 Ablation and Case Study

**Component Ablation.** We assess Attention Column Formation (ACF), Visual Summary Absorption (VSA) Semantic Target Alignment (STA), and Sink Semantic Condition (SSC), designing SISA variants to test performances. As shown in Table 4, all components contribute to the attacks.

**Visualization.** We compare the sink-regime behavior induced by SISA at position $j^\star$ against the *Clean* baseline by similarly adopting the columnar-attention strength $\mathrm{ACF}^\ell(j^\star) = \frac{1}{|\mathcal{I}_{\to}(j^\star)|}\sum_{i\in\mathcal{I}_{\to}(j^\star)} A_{i,j^\star}^\ell$ and the visual-summary resemblance $\mathrm{VSA}^\ell(j^\star) = \cos\big(h_{j^\star}^\ell, h_g^\ell\big)$. Unless stated otherwise, all results are reported on MLLMU-Bench (case(i)), where all victim models are unlearned using SIU. As shown in Figure 5, SISA consistently increases both $\mathrm{ACF}^\ell(j^\star)$ and $\mathrm{VSA}^\ell(j^\star)$ across victim architectures, and the paired frequency $\Pr[\text{SISA} \ge \textit{Clean}]$ is high for both metrics.

## 4.5 Results against Countermeasures

We evaluate three potential countermeasures: ***(i) CIDER*** [Xu *et al.*, 2024] detects malicious inputs by measuring the embedding-space image–text alignment shift before and after

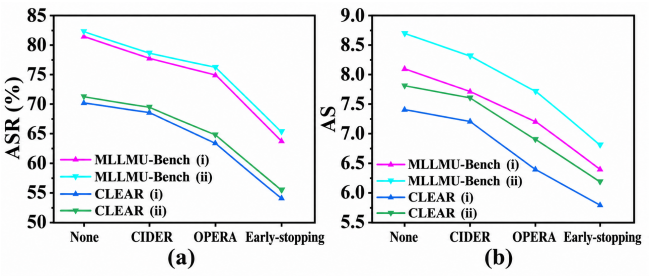


Figure 6: Defense evaluation against SISA. (a) ASR(%) and (b) AS.

denoising, and returns a refusal upon detection. ***(ii) OPERA*** [Huang *et al.*, 2024] is a decoding-time mitigation designed to suppress attention-sink/columnar behaviors by introducing sink-aware logit penalties during beam-search decoding. ***(iii) Early-stopping*** [Wang *et al.*, 2025] is an adaptive mitigation that detects the attention-sink phenomenon during generation and terminates output before sink tokens appear.

**Results:** Figure 6 shows that CIDER modestly mitigates SISA. This is expected because SISA constrains perturbations within the budget and promotes semantic consistency rather than image–text misalignment, which reduces the shift used for detection. OPERA attenuates SISA by suppressing columnar behaviors. Nevertheless, SISA remains effective, as semantic anchoring still provides guidance towards the desired output beyond sink-regime induction, with ASR and AS above 75.8% and 7.7 on MLLMU-Bench. Early-stopping reduces the incidence of elicited target, yet it increases perplexity (adapted from [Wang *et al.*, 2025]) by 45.36% on average.

## 5 Related Work

**Adversarial Inputs in VLMs.** Recent work has shown that VLMs can be manipulated by adversarial image–text inputs. Jailbreak-style methods primarily disguise malicious intent and elicit non-predefined responses by bypassing refusal behaviors. Typography-based attacks such as FigStep [Gong *et al.*, 2025] embed structured yet incomplete textual patterns into images and prompt the model to continue them, weakening safety-aligned refusals. Optimization-based methods iteratively tune inputs to induce harmful responses. For example, VAJM [Qi *et al.*, 2024] shows that a single optimized image can jailbreak multiple prompts, while UMK [Wang *et al.*, 2024] jointly optimizes adversarial text–image pairs to improve robustness across prompts and models.

A concurrent line of work studies transferable attacks that steer VLM behaviors via targeted feature alignment. These methods optimize perturbations such that the perturbed input is encoded closer to a chosen target in representation space (e.g., concept/image/text embedding). For instance, AnyAttack [Zhang *et al.*, 2025a] leverages large-scale self-supervised pretraining for transferability, M-Attack [Li *et al.*, 2025] improves transfer via random cropping and resizing during optimization, and FOA-Attack [Jia *et al.*, 2025] jointly aligns global and local patch-token features via clustering.

Despite these advances, existing VLM attacks are not specially tailored to target elicitation under VLM unlearning, which is essential for evaluating the reliability of expected unlearning behavior [Yuan *et al.*, 2025; Schwinn *et al.*, 2024].

**Adversarial Inputs in VLM Unlearning.** There is limited literature on adversarial attacks against unlearned VLMs. SUA was proposed to craft perturbations on the visual modality that trigger unlearned VLMs to reveal supposedly forgotten content [Zhang *et al.*, 2025b]. SUA assumes white-box access to the victim unlearned VLM (e.g., architecture and parameters) for gradient-based optimization, or gray-box access (e.g., output-logit queries) for zeroth-order updates. However, unlearned VLMs are often proprietary, making such access unavailable or repeated queries suspicious.

Recent research has identified a transformer-rooted phenomenon across VLMs, known as visual sink tokens, where certain visual tokens attract disproportionate attention during generation and behave as low-semantic, summary-like visual aggregators [Kang *et al.*, 2025; Luo *et al.*, 2025]. Leveraging the generalizability of sink-token formation across VLMs, recent work explores transferable hallucination attacks [Wang *et al.*, 2025]. In this paper, we propose SISA, a surrogate-only attack framework exploiting the visual sink token. SISA crafts transferable adversarial inputs by inducing a sink token as a structural anchor and enforcing sink-conditioned semantic alignment for targeted elicitation.

## 6 Conclusion

In this paper, we propose SISA, the first adversarial attack framework against VLM unlearning under a more relaxed access setting. Unlike prior attacks, SISA does not rely on the unlearned model's architecture, parameters, or output-logit queries. SISA induces a sink-like behavior at a visual token and aligns model output to the target while conditioning attention on the induced sink token. The persistence of sink token and sink-conditioned semantic alignment enable crafting transferable adversarial inputs, thereby undermining the reliability of expected forgetting behavior. Extensive experiments demonstrate that SISA achieves notable performance across settings and outperforms state-of-the-art attacks.

## Acknowledgments

This work was supported by NSFC (Grant No. U2241213), Guangxi Key Research and Development Project (Guike AB25069120), NSFC (Grant No. 62272038).

## References

[Andriushchenko *et al.*, 2023] Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *ICML*, volume 202, pages 840–902, 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, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[Barbero *et al.*, 2025] Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do llms attend to the first token? *arXiv preprint arXiv:2504.02732*, 2025.

[Cui *et al.*, 2024] Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *CVPR*, pages 24625–24634, 2024.

[Dai *et al.*, 2023] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

[Darcet *et al.*, 2023] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *ICLR*, 2023.

[Deitke *et al.*, 2025] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *CVPR*, pages 91–104, 2025.

[Dontsov *et al.*, 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 ACL 2025*, pages 20582–20603, 2025.

[Doshi and Stickland, 2024] Jai Doshi and Asa Cooper Stickland. Does unlearning truly unlearn? a black box evaluation of llm unlearning methods. *arXiv preprint arXiv:2411.12103*, 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 *AAAI*, volume 39, pages 23951–23959, 2025.

[Gu *et al.*, 2025] Tianle Gu, Kexin Huang, Ruilin Luo, Yuanqi Yao, Xiuying Chen, Yujiu Yang, Yan Teng, and Yingchun Wang. From evasion to concealment: Stealthy knowledge unlearning for llms. In *Findings of ACL 2025*, pages 10261–10279, 2025.

[Hong *et al.*, 2025] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

[Huang *et al.*, 2024] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*, pages 13418–13427, 2024.

[Huo *et al.*, 2025] 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. *arXiv preprint arXiv:2502.11051*, 2025.

[Jia *et al.*, 2025] Xiaojun Jia, Sensen Gao, Simeng Qin, Tianyu Pang, Chao Du, Yihao Huang, Xinfeng Li, Yiming Li, Bo Li, and Yang Liu. Adversarial attacks against closed-source mllms via feature optimal alignment. *arXiv preprint arXiv:2505.21494*, 2025.

[Kang *et al.*, 2025] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*, 2025.

[Keskar *et al.*, 2016] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[Li *et al.*, 2022] Bo Li, Yongqiang Yao, Jingru Tan, Gang Zhang, Fengwei Yu, Jianwei Lu, and Ye Luo. Equalized focal loss for dense long-tailed object detection. In *CVPR*, pages 6990–6999, 2022.

[Li *et al.*, 2024] Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. Single image unlearning: Efficient machine unlearning in multimodal large language models. *NeurIPS*, 37:35414–35453, 2024.

[Li *et al.*, 2025] Zhaoyi Li, Xiaohan Zhao, Dong-Dong Wu, Jiacheng Cui, and Zhiqiang Shen. A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of gpt-4.5/4o/o1. *arXiv preprint arXiv:2503.10635*, 2025.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023.

[Liu *et al.*, 2025a] 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 *NAACL-HLT 2025 (Volume 1: Long Papers)*, pages 4105–4135, 2025.

[Liu *et al.*, 2025b] Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, and Meng Jiang. Modality-aware neuron pruning for unlearning in multimodal large language models. *arXiv preprint arXiv:2502.15910*, 2025.

[Luo *et al.*, 2025] Jiayun Luo, Wan-Cyuan Fan, Lyuyang Wang, Xiangteng He, Tanzila Rahman, Purang Abolmaesumi, and Leonid Sigal. To sink or not to sink: Visual information pathways in large vision-language models. *arXiv preprint arXiv:2510.08510*, 2025.

[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. *arXiv preprint arXiv:2401.06121*, 2024.

[Patil *et al.*, 2024] 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. *Trans. Mach. Learn. Res.*, 2024.

[Qi *et al.*, 2024] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *AAAI*, volume 38, pages 21527–21536, 2024.

[Rimsky *et al.*, 2024] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *ACL 2024*, pages 15504–15522, 2024.

[Schwinn *et al.*, 2024] Leo Schwinn, David Dobre, Sophie Xhonneux, Gauthier Gidel, and Stephan Günnemann. Soft prompt threats: Attacking safety alignment and unlearning in open-source llms through the embedding space. *NeurIPS*, 37:9086–9116, 2024.

[Sun *et al.*, 2024] Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

[Team *et al.*, 2025] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[Wang *et al.*, 2023] Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*, 2023.

[Wang *et al.*, 2024] Ruofan Wang, Xingjun Ma, Hanxu Zhou, Chuanjun Ji, Guangnan Ye, and Yu-Gang Jiang. White-box multimodal jailbreaks against large vision-language models. In *MM '24*, pages 6920–6928, 2024.

[Wang *et al.*, 2025] Yining Wang, Mi Zhang, Junjie Sun, Chenyue Wang, Min Yang, Hui Xue, Jialing Tao, Ranjie Duan, and Jiexi Liu. Mirage in the eyes: Hallucination attack on multi-modal large language models with only attention sink. In Lujo Bauer and Giancarlo Pellegrino, editors, *USENIX Security 2025*, pages 3707–3726, 2025.

[Willemsen *et al.*, 2023] Bram Willemsen, Livia Qian, and Gabriel Skantze. Resolving references in visually-grounded dialogue via text generation. In *SIGDIAL 2023*, pages 457–469, 2023.

[Xiao *et al.*, 2023] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

[Xu *et al.*, 2024] Yue Xu, Xiuyuan Qi, Zhan Qin, and Wenjie Wang. Cross-modality information check for detecting jailbreaking in multimodal large language models. *arXiv preprint arXiv:2407.21659*, 2024.

[Yao *et al.*, 2024] Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. Machine unlearning of pre-trained large language models. *arXiv preprint arXiv:2402.15159*, 2024.

[Yuan *et al.*, 2025] Hongbang Yuan, Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Towards robust knowledge unlearning: An adversarial framework for assessing and improving unlearning robustness in large language models. In *AAAI*, volume 39, pages 25769–25777, 2025.

[Zhang *et al.*, 2025a] Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *CVPR*, pages 19900–19909, 2025.

[Zhang *et al.*, 2025b] Xianren Zhang, Hui Liu, Delvin Ce Zhang, Xianfeng Tang, Qi He, Dongwon Lee, and Suhang Wang. Sua: Stealthy multimodal large language model unlearning attack. In *EMNLP 2025*, page 11230–11243, 2025.

[Zhu *et al.*, 2025] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[Zou *et al.*, 2025] Xingchen Zou, Yibo Yan, Xixuan Hao, Yuehong Hu, Haomin Wen, Erdong Liu, Junbo Zhang, Yong Li, Tianrui Li, Yu Zheng, et al. Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. *Inf. Fusion*, 113:102606, 2025.