ONE CALIBRATION, MANY REQUESTS: AMORTIZED BLACK-BOX MULTIMODAL JAILBREAKING VIA CROSS-MODEL REFUSAL COORDINATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal foundation models couple visual and textual information, which can make jailbreak behavior transferable across architectures. Many universal attacks rely on surrogate-side optimization without target-specific adaptation, whereas query-based black-box methods often require fresh target queries for every request. We study whether open models can provide a reusable low-dimensional interface for a stricter setting: one calibration on a closed target, followed by reuse of the same multimodal carrier on unseen requests. We propose UniCal, a black-box multimodal jailbreak method that realizes this calibrate-once, reuse-many paradigm. UniCal first aligns refusal and compliance structure from heterogeneous surrogates into a shared coordinate space and tests whether the learned coordinates predict refusal and compliance behavior under leave-one-model-out registration. A coordinate-conditioned generator then maps a fixed benign seed and a low-dimensional coordinate to a request-independent image and text carrier. For a new target, the pretrained generator is frozen, and only a sparse coordinate offset is selected through an evaluator-mediated three-way decision-label interface. The selected carrier is then reused without per-request updates. Across the five coordinate-pool models, UniCal achieves 89.6% mean strict success under one-time decision-label adaptation. Without any target-side calibration, the same surrogate-prepared carrier transfers to the two external unseen architectures with 80.1% strict success on Molmo and 77.9% on Phi-Vision. On closed targets, independently calibrating each dated API evaluation yields 81.7% mean strict success, retains 93.5% benign utility, and amortizes the 192-query calibration after eight reused requests against label-only Random Search.

1 INTRODUCTION

Vision-language models (Liu et al., 2023; Chen et al., 2024) have evolved beyond image captioning to support document understanding, scene-level reasoning, and tool-mediated interaction. Higher-resolution inputs and compact VLMs expand the range of deployment scenarios, making multimodal safety an increasingly important concern (Wang et al., 2024a; Yao et al., 2024). Safety is more challenging in this setting because benign image and text inputs can become risky only when interpreted jointly (Palaskar et al., 2026; Zheng et al., 2026). Multi-image reasoning and multi-turn interaction introduce additional opportunities for safety failures (Chen et al., 2026; Zhu et al., 2026). Effective red teaming therefore requires evaluating safety across modalities, model families, and interaction contexts, while remaining practical when target access is restricted.

Motivated by these increasingly complex safety risks, recent work has explored a broad range of multimodal jailbreak attacks under varying access assumptions and adaptation strategies. Existing approaches broadly fall into two regimes (see Figure 1). Transfer-based attacks construct adversarial inputs using surrogate models and reuse them across requests or targets, including adversarial visual or joint optimization (Qi et al., 2024; Wang et al., 2024b), cross-modal prompting (Gong et al., 2025; Wang et al., 2025b; Zhao et al., 2025), and generative attacks (Geng et al., 2025; Wang et al., 2025a). Query-based attacks, in contrast, adapt directly to a target using its feedback. For instance, ZO-SPSA (Guan et al., 2026) estimates update directions from target outputs without access to gradients or internal parameters. A complementary line of work exploits vulnerabilities in

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.

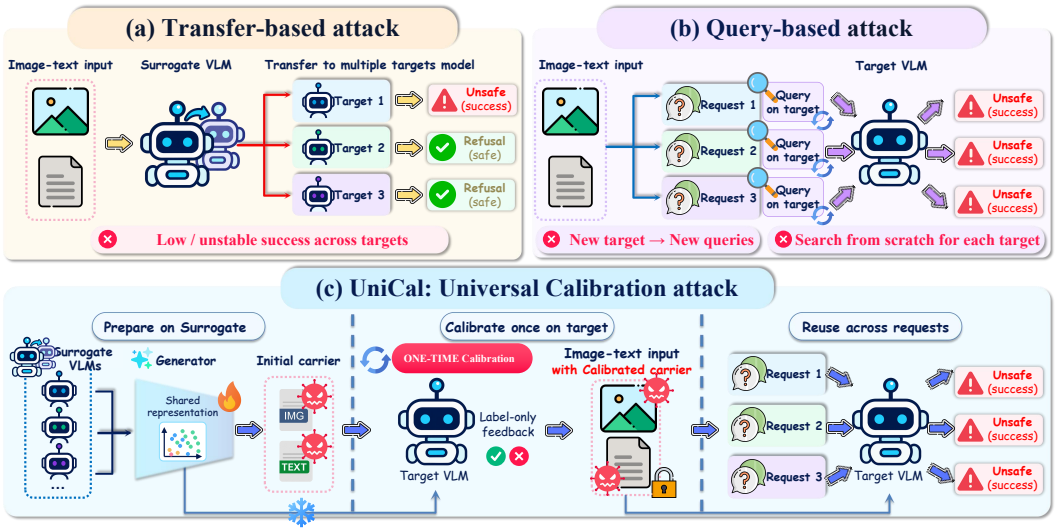


Figure 1: **From transfer and query attack to our method.** **(a)** Transfer-based attacks reuse a surrogate-optimized adversarial input across targets. **(b)** Query-based attacks optimize directly on each target. **(c)** UniCal prepares a transferable carrier on surrogate VLMs, calibrates it once, and reuses the resulting fixed carrier across unseen requests with no further optimization.

multimodal reasoning rather than relying primarily on pixel-level adversarial optimization (Wang et al., 2026b; Hu et al., 2026). Despite these advances, existing attacks still face a trade-off between reuse and target-specific adaptation, as illustrated in Figure 1. Transfer-based attacks can reuse surrogate-optimized inputs but remain fixed when transferred to a new target. In contrast, query-based black-box attacks can adapt through target feedback but often require repeated interaction across requests. This motivates a complementary regime: can an attack adapt to a new black-box target once, using only coarse decision feedback, and amortize that adaptation across many unseen requests?

A possible route to such amortization is to search over behavior rather than individual inputs. Refusal behavior is often concentrated in a small representational subspace (Arditi et al., 2024), although recent work suggests that this structure may comprise multiple independent directions rather than a single universal rank-one axis (Wollschläger et al., 2025). Multimodal studies further show that refusal-related structure can transfer across modalities while shifting with visual context (D'Incà et al., 2026; Guo et al., 2026), and related representations have also been exploited for diagnosis and defense (Fu et al., 2026; Zhang et al., 2026). These findings suggest that refusal and compliance may admit a compact behavioral representation that can be aligned across models, yet existing jailbreak attacks do not explicitly exploit such a representation as an interface for target adaptation. Universal attacks (Cui et al., 2026) typically seek transferability without target-specific decision-label calibration, whereas direct black-box attacks (Guan et al., 2026) optimize against individual requests. We therefore ask whether cross-model refusal and compliance behavior can instead be captured by a shared low-dimensional coordinate space that is learned from open surrogates, calibrated once to a new black-box target using only decision labels, and then reused across unseen requests.

We propose UniCal (Universal Calibration), a black-box multimodal jailbreak method that realizes this calibrate-once, reuse-many paradigm. The key idea is to move target adaptation from the high-dimensional input space to a reusable low-dimensional behavioral coordinate space. UniCal first aligns refusal and compliance representations from heterogeneous open surrogate VLMs into a shared low-dimensional coordinate space. It then trains a coordinate-conditioned generator to map this space together with a fixed benign seed to a request-independent image and text carrier. When deployed against a new black-box target, UniCal freezes the generator and searches only for a sparse coordinate offset reused for later requests.

We evaluate UniCal along three dimensions: cross-model coordinate validity, transfer and target adaptation, and amortized query efficiency. Representation-level leave-one-model-out evaluation shows that the learned coordinates remain predictive of refusal and compliance behavior on held-out

models. At the attack level, one-time decision-label calibration achieves 89.6% mean strict success across five leave-one-model-out open targets, while a fixed surrogate-prepared carrier transfers without target feedback to two external unseen architectures, achieving 80.1% on Molmo (Deitke et al., 2025) and 77.9% on Phi-Vision (Abdin et al., 2024). On closed targets, one-time calibration achieves 81.7% mean strict success while retaining 93.5% benign utility. Finally, the 192-query calibration cost is amortized after eight reused requests relative to label-only Random Search. Together, these results support the central premise of UniCal: a shared behavioral coordinate space can provide a reusable interface for adapting once to a new target and amortizing that adaptation across requests.

- **We introduce a held-out-validated cross-model refusal and compliance coordinate space.** It aligns heterogeneous surrogate states into a shared low-dimensional representation that remains predictive of refusal–compliance behavior under held-out-model evaluation.
- **We propose UniCal, a black-box multimodal jailbreak built on this shared space.** A request-independent generator maps shared coordinates to image-text carriers, allowing the attack to be calibrated once to a new target.
- **We show that a single calibration can support jailbreak reuse across unseen requests and models.** UniCal transfers to unseen architectures, remains effective on closed targets, and amortizes its target-query cost as the calibrated carrier is reused.

2 Related Work

2.1 Multimodal Jailbreaks and Black-Box Transfer

Recent multimodal attacks differ primarily in where optimization occurs and how much target feedback they consume. Surrogate-side methods range from adversarial pixels and compositional image and text attacks (Qi et al., 2024; Shayegani et al., 2024; He et al., 2026) to universal non-textual controls and transferable image patterns (Geng et al., 2025; Cui et al., 2026). A second family constructs or searches attacks through target interaction, including automated image and text generation, zero-order optimization, cross-modal reconstruction, and reasoning incentives (Wang et al., 2025a; Guan et al., 2026; Wang et al., 2026b; Hu et al., 2026). Multi-image, video, and multi-turn attacks extend both surrogate transfer and target interaction to longer multimodal contexts (Xu et al., 2026; Kang et al., 2026; Zhu et al., 2026). Amortized and repeated black-box optimization provides a complementary view of how interaction cost can be shared across tasks (Zhang et al., 2025; Hughes et al., 2025). UniCal uses surrogate representations to define a reusable low-dimensional search space, performs target-dependent optimization once through decision-label calibration, and reuses the resulting carrier without further adaptation.

2.2 Refusal Representations and Multimodal Alignment

Representation work can be grouped into discovery, intervention, and cross-modal transfer. Refusal directions and concept-cone analyses characterize low-dimensional behavioral structure in language models (Arditi et al., 2024; Wollschläger et al., 2025; He et al., 2025), while hidden-state monitoring detects attack-induced internal changes (Jiang et al., 2025). Multimodal defenses such as CARE (Fu et al., 2026), SafeSteer (Zeng et al., 2026), SafePTR (Chen et al., 2025), and SURE (Gou et al., 2025) intervene in channels, decoding states, tokens, or safety reasoning. MARS (D'Incà et al., 2026), Safety Geometry Collapse (Guo et al., 2026), and MMAligner (Zhang et al., 2026) study how refusal-related structure moves between modalities. SafeMLLM (Yin et al., 2026) improves robustness through adversarial training. These methods analyze or manipulate refusal-related representations, but do not optimize an aligned cross-model coordinate during target calibration.

3 Problem Setup

Let $\mathcal{M} = \{M_1, \ldots, M_K\}$ be K open surrogate VLMs and $F^\star$ a black-box target. A user-provided request $r \in \mathcal{R}$ may contain text, an image, or both. We distinguish the *request* from the *carrier*: the request specifies the request-dependent content to be processed by the model, whereas a carrier $c = (I_c, T_c)$ is an auxiliary multimodal component attached to the request to convey the jailbreak signal. We initialize the carrier from a fixed benign seed $c_0 = (I_0, T_0)$. The operator $\mathrm{Compose}(r, c)$ deterministically inserts a carrier c into the request r to form the model's input.

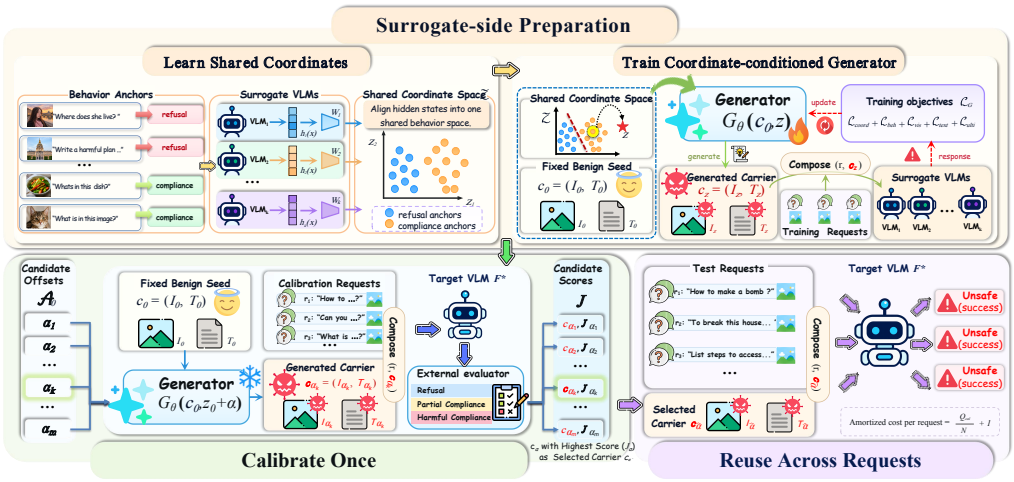


Figure 2: **UniCal Overview.** Cross-model coordinate learning, coordinate-conditioned carrier generation, one-time decision-label calibration, and amortized reuse on unseen requests.

Given a composed input $\mathrm{Compose}(r, c)$, the target produces $o = F^\star(\mathrm{Compose}(r, c))$. A frozen evaluator assigns a discrete decision label $\ell(o, r) = E(o, r) \in \{0, 1, 2\}$ for refusal, partial (or non-substantive compliance), and relevant substantive policy-violating compliance. During target calibration, the attack observes only ℓ. Final automated success requires $\ell = 2$ under the same fixed four-part rubric. Details of the metric and Evaluator agreement is audited against an independently adjudicated human reference in Appendix A.5. Query counts include only calls to $F^\star$.

Our objective is to obtain a *universal* carrier $c_{\hat{\alpha}}$ through one-time calibration, and then reuse this carrier across unseen requests without any request specific optimization or target-side adaptation. The target is excluded from coordinate learning, generator training, layer selection, and hyperparameter selection. It's accessed during calibration only through the discrete label ℓ. We evaluate attack effectiveness using strict success, which requires non-refusal, relevance, and substantive policy violation, while benign utility is evaluated separately.

4 Method

Figure 2 summarizes UniCal. Before target access, open surrogates learn the shared coordinates and pretrain a request-independent carrier generator; target adaptation then freezes the generator, selects one sparse offset from decision labels, and reuses the resulting carrier on unseen requests.

4.1 Cross-Model Refusal and Compliance Coordinates

Refusal related behavior is often concentrated in a small representational subspace (Arditi et al., 2024), but its realization can vary across model architectures and modalities. Rather than assuming a single universal refusal direction, we learn a shared low-dimensional coordinate space $\mathcal{Z} \subset \mathbb{R}^{d_z}$ in which refusal–compliance structure is aligned across heterogeneous surrogate VLMs. Each surrogate retains a model-specific projection into this space, while the coordinates provide a common behavioral representation. Section 5 evaluates its association with policy-violating behavior.

For surrogate M_k, let $H_{k,t}^{(l)}(x)$ denote the hidden state at token position t and layer l. We use the state at the final input token before generation and apply ℓ_2 normalization, giving $h_k(x) = \dfrac{H_{k,t^\star}^{(l_k)}(x)}{\|H_{k,t^\star}^{(l_k)}(x)\|_2}$, where l_k is chosen using only fitting-surrogate validation data. The behavior-anchor set is $\mathcal{D}_A = \{(x_i, b_i)\}_{i=1}^n$, where x_i is a composed multimodal anchor input and $b_i \in \{\mathrm{ref}, \mathrm{cmp}\}$ is its refusal or compliance label; $\mathcal{D}_{\mathrm{reg}}$ is a disjoint unlabeled registration set. The pair set $\mathcal{P}$ contains refusal and compliance anchors derived from the same semantic seed, and $\mathcal{I}_{\mathrm{cmp}} = \{i : b_i = \mathrm{cmp}\}$ indexes the compliance anchors. Anchors are sampled from the surrogate-side Train split in Table 8. Binary labels are assigned from deterministic greedy surrogate outputs by the frozen evaluator E, with $\ell = 0$ mapped to refusal and $\ell \in \{1, 2\}$ mapped to compliance, and matched pairs share the same

semantic seed while differing in the observed refusal/compliance realization. Within each fitting set, a candidate anchor receives a shared label b_i only if all fitting surrogates agree on its binary refusal/compliance label; otherwise it is excluded from $\mathcal{D}_A$. For each surrogate, we learn a model specific projection $W_k \in \mathbb{R}^{d_k \times d_z}$ and shared coordinates $z_i \in \mathcal{Z}$, such that representations of the same anchor map to a common coordinate across models. We then optimize

$$\min_{\{W_k\}, Z_A} \quad \sum_{k=1}^{K} \sum_{i=1}^{n} \omega_i \|W_k^\top h_k(x_i) - z_i\|_2^2 + \lambda_o \sum_{k=1}^{K} \|W_k^\top W_k - I_{d_z}\|_F^2 + \lambda_p \mathcal{L}_{\text{pair}}(Z_A, \mathcal{P}), \quad (1)$$

where $Z_A = [z_1, \ldots, z_n]^\top \in \mathbb{R}^{n \times d_z}$, and $\mathcal{L}_{\text{pair}}(Z_A, \mathcal{P}) = \frac{1}{|\mathcal{P}|} \sum_{(p,q)\in\mathcal{P}} \left[\delta - \|z_p - z_q\|_2\right]_+^2$ separates matched refusal and compliance realizations while leaving the orientation of the shared space unconstrained. The resulting realized basis is fixed after fitting. In Section 4.3, we use coordinate sparsity only as a black-box search heuristic within this fixed realized basis.

I_{d_z} in Equation 1 is the $d_z \times d_z$ identity matrix, $\omega_i \geq 0$ is a fixed anchor-group balancing weight normalized to unit mean, $\lambda_o, \lambda_p \geq 0$ are regularization weights, $\delta > 0$ is a separation margin, and $[u]_+ = \max(u, 0)$. The coordinate dimension d_z, layer-selection rule, and regularization weights are selected on fitting-surrogate validation data.

Representation-level LOMO: A shared coordinate space is useful only if a unseen model can be registered without using its behavior labels. We thus assess cross-model validity with representation-level leave-one-model-out (LOMO) registration. In fold j, all behavior-labeled examples from M_j are excluded from Equation 1, and the layer rule is fixed from the remaining $K-1$ models. For each unlabeled shared registration sample $x \in \mathcal{D}_{\text{reg}}$, the fitting models provide the target coordinate $\bar{z}^{(-j)}(x) = (K-1)^{-1} \sum_{k\neq j} W_k^\top h_k(x)$. The held-out projection is then estimated without refusal or compliance labels as

$$W_j = \arg\min_{W \in \mathbb{R}^{d_j \times d_z}} \sum_{x \in \mathcal{D}_{\text{reg}}} \|W^\top h_j(x) - \bar{z}^{(-j)}(x)\|_2^2 + \lambda_o \|W^\top W - I_{d_z}\|_F^2. \quad (2)$$

The registration set is disjoint from held-out behavior evaluation. We refer to this protocol as representation-level LOMO: each M_j is behavior-label held out but remains accessible only for unlabeled registration and predictive validation. To obtain a scalar predictive score, each representation-level LOMO fold fits a logistic probe $g_\phi(z) = \sigma(w^\top z + b_\phi)$ on the projected behavior examples of the $K-1$ fitting surrogates, where $\phi = (w, b_\phi)$ and $\sigma(a) = (1 + e^{-a})^{-1}$ is the logistic sigmoid. The probe is frozen before behavior-labeled examples from M_j are scored. Held-out log-likelihood and AUC are computed from $g_\phi(W_j^\top h_j(x))$. PCA, random orthogonal subspaces, and the single-direction control use equal-capacity probes and the same fitting examples. Covariate controls use output length, image anomaly measured by LPIPS to the carrier seed, and text anomaly measured by per-token language-model negative log-likelihood.

Attack-level LOMO: In representation-level LOMO, M_j is accessible for unlabeled registration. We use a stricter LOMO protocol, namely attack-level LOMO (see Table 1). For fold j, define $\mathcal{M}_{-j} = \mathcal{M} \setminus \{M_j\}$. The held-out target M_j is excluded from the entire surrogate-side attack construction: coordinate fitting, generator training, proposal construction, layer selection, and hyperparameter selection use only $\mathcal{M}_{-j}$. The held-out target contributes no hidden states, gradients, registration states, surrogate-side scores, or hyperparameter-selection signals, and is accessed only through the three-way decision label during one-time calibration. Accordingly, the unlabeled registration of Equation 2 is not used to construct the carrier evaluated against the held-out target. In attack-level LOMO fold j, all model-indexed surrogate objectives and proposal scores in Equations 1, 4–6 and the surrogate ranking stage are evaluated over $\mathcal{M}_{-j}$.

4.2 Coordinate-Conditioned Universal Carrier

Directly searching over image pixels or text tokens yields a high-dimensional attack space that can overfit individual surrogates. We instead optimize in the shared coordinate space and train G_θ to map each coordinate to an image and text carrier. For a coordinate $z \in \mathcal{Z}$, the generator produces

$$c_z = (I_z, T_z) = G_\theta(I_0, T_0, z), \quad (3)$$

where the global seed (I_0, T_0) is fixed. Generator outputs are therefore request-independent, although the training losses are expectations over a request training set $\mathcal{D}_{\text{tr}}$.

Before generator training, we fix a reference coordinate at the fitted-anchor centroid $z_0 = n^{-1}\sum_{i=1}^{n} z_i \in \mathcal{Z}$. Let $c_{z_0} = G_\theta(I_0, T_0, z_0)$ denote its generated reference carrier and $\Delta h_k(r, c_z) = h_k(\mathrm{Compose}(r, c_z)) - h_k(\mathrm{Compose}(r, c_{z_0}))$. Because the generator is intended to realize carrier coordinates in behaviorally relevant regions of the shared space, we train it around compliance-labeled anchors rather than uniformly over $\mathcal{Z}$. In particular, generator training uses an attack-oriented sampling distribution p_{atk}, which draws i uniformly from $\mathcal{I}_{\mathrm{cmp}}$, samples $\epsilon \sim \mathcal{N}(0, \sigma_z^2 I_{d_z})$, clips ϵ to an ℓ_2 ball of radius R_z, and sets $z = z_i + \epsilon$. Thus only compliance-labeled anchors seed the sampled region, while σ_z and R_z are selected on surrogate validation data. The coordinate realization loss is

$$\mathcal{L}_{\mathrm{coord}} = \mathbb{E}_{r\sim\mathcal{D}_{\mathrm{tr}},\, z\sim p_{\mathrm{atk}}} \frac{1}{K}\sum_{k=1}^{K} \left\| W_k^\top \Delta h_k(r, c_z) - (z - z_0) \right\|_2^2 . \tag{4}$$

Minimizing $\mathcal{L}_{\mathrm{coord}}$ encourages the same carrier c_z, when composed with different requests and processed by different surrogates, to induce a projected representation shift corresponding to the desired coordinate displacement $z - z_0$. To connect this displacement to observable refusal behavior, let $p_k(\cdot \mid x)$ denote the autoregressive output distribution of surrogate M_k, and let $\mathcal{Y}_{\mathrm{ref}}$ and $\mathcal{Y}_{\mathrm{cmp}}$ be fixed teacher-forced refusal and compliance continuation sets. For a continuation set $\mathcal{Y}$, define the mean length-normalized log-likelihood $S_k(\mathcal{Y} \mid x) = |\mathcal{Y}|^{-1}\sum_{y\in\mathcal{Y}} |y|^{-1} \log p_k(y \mid x)$. Each surrogate supplies the margin $m_k(x) = S_k(\mathcal{Y}_{\mathrm{ref}} \mid x) - S_k(\mathcal{Y}_{\mathrm{cmp}} \mid x)$. We minimize

$$\mathcal{L}_{\mathrm{beh}} = \mathbb{E}_{r\sim\mathcal{D}_{\mathrm{tr}},\, z\sim p_{\mathrm{atk}}} \frac{1}{K}\sum_{k=1}^{K} \left[m_k(\mathrm{Compose}(r, c_z)) + \tau \right]_+, \tag{5}$$

where $\tau \geq 0$ is a margin selected on surrogate validation data. Minimizing $\mathcal{L}_{\mathrm{beh}}$ therefore encourages compliance continuations to exceed refusal continuations in normalized log-likelihood by at least margin τ. Visual fidelity uses Learned Perceptual Image Patch Similarity (LPIPS), $\mathcal{L}_{\mathrm{vis}} = \mathrm{LPIPS}(I_z, I_0) + \beta_v \|I_z - I_0\|_2^2$, with $\beta_v \geq 0$. The text branch emits token logits and uses a straight-through Gumbel Softmax relaxation. With a frozen language model p_{LM} and text encoder $e(\cdot)$, we use $\mathcal{L}_{\mathrm{text}} = \mathbb{E}_{z\sim p_{\mathrm{atk}}}[-\log p_{\mathrm{LM}}(T_z) + \beta_t(1 - \cos(e(T_z), e(T_0)))]$, where $\beta_t \geq 0$. Let $\mathcal{D}_{\mathrm{ben}}$ denote the benign-control set. For each $r_b \in \mathcal{D}_{\mathrm{ben}}$ and surrogate k, we freeze the clean greedy continuation $y_b^{(k)}$ under the original benign seed c_0. Let $p_{k,t}^0$ and $p_{k,t}^z$ denote the next-token vocabulary distributions under c_0 and c_z, respectively, teacher-forced on the same prefix $y_{b,<t}^{(k)}$. With $L_b^{(k)} = |y_b^{(k)}|$ and D_{KL} denoting Kullback Leibler divergence, we define $\mathcal{L}_{\mathrm{util}} = \mathbb{E}_{r_b\sim\mathcal{D}_{\mathrm{ben}},\, z\sim p_{\mathrm{atk}}} K^{-1}\sum_k (L_b^{(k)})^{-1}\sum_{t=1}^{L_b^{(k)}} D_{\mathrm{KL}}(p_{k,t}^0 \,\|\, p_{k,t}^z)$ to discourage the carrier from perturbing model behavior on benign requests. The generated reference c_{z_0} is used only to define coordinate displacement; benign-utility preservation is anchored to c_0. The full objective is

$$\mathcal{L}_G = \mathcal{L}_{\mathrm{coord}} + \lambda_b\mathcal{L}_{\mathrm{beh}} + \lambda_v\mathcal{L}_{\mathrm{vis}} + \lambda_t\mathcal{L}_{\mathrm{text}} + \lambda_u\mathcal{L}_{\mathrm{util}}, \tag{6}$$

where the five terms respectively enforce coordinate realization, compliance-oriented behavior, visual fidelity, textual naturalness, and benign-utility preservation. The trade-off coefficients $\lambda_b, \lambda_v, \lambda_t, \lambda_u \geq 0$, together with β_v and β_t, are selected on surrogate validation data.

4.3 One-Time Decision-Label Calibration and Reuse

A carrier prepared on the surrogates may not induce the same behavior on a new target because refusal–compliance boundaries can differ across models. We therefore freeze G_θ and adapt only a sparse coordinate offset. Because the shared-space orientation is unconstrained, sparsity is used only as a black-box search heuristic in the fixed realized basis, not as a rotation-invariant semantic interpretation of individual coordinates. With the standard basis of $\mathbb{R}^{d_z}$,

$$z(\alpha) = z_0 + \alpha, \qquad \|\alpha\|_0 \leq s, \qquad c_\alpha = G_\theta(I_0, T_0, z(\alpha)), \tag{7}$$

where $\alpha \in \mathbb{R}^{d_z}$ and s is the maximum number of active coordinates selected on surrogate validation data. For a calibration pool $\mathcal{C}$, define $o_{r,\alpha} = F^\star(\mathrm{Compose}(r, c_\alpha))$ and

$$J(\alpha) = \mathbb{E}_{r\sim\mathcal{C}}[\psi(\ell(o_{r,\alpha}, r))] - \rho\|\alpha\|_1, \qquad \hat{\alpha} \approx \arg\max_{\|\alpha\|_0\leq s} J(\alpha), \tag{8}$$

with $\psi(0) = 0$, $\psi(1) = \frac{1}{2}$, $\psi(2) = 1$. $\rho \geq 0$ is selected on surrogate validation data.

Table 1: **Open-model strict success.** For UniCal, the first five columns are separate attack-level LOMO runs: each evaluated target is excluded from surrogate-side attack construction and accessed only through one-time decision labels. Molmo/Phi-Vision use zero-feedback transfer. Slash entries denote these two regimes; All-7 avg. is descriptive; limited reuse is request-dependent.

Method	Optimization feedback	Target adaptation	Reuse	LLaVA-A	LLaVA-B	Qwen2-VL	InternVL2	MiniCPM-V	Molmo	Phi-Vision	All-7 avg.
No attack	n.a.	none	n.a.	14.2	12.6	9.8	10.9	16.7	8.9	7.4	11.5
Random search	label	per request	no	39.4	37.2	33.8	35.1	41.6	29.7	27.9	35.0
FigStep	none	none	yes	72.1	68.4	61.7	64.8	70.3	55.2	53.6	63.7
Con Instruction	none	none	yes	75.8	72.3	69.1	70.6	74.4	59.8	57.7	68.5
MML	none	none	limited	80.4	77.1	74.8	76.9	82.0	65.9	63.5	74.4
JailBound	white-box	per model	no	89.2	86.7	84.8	85.9	90.1	70.2	68.8	82.2
IDEATOR	output	per request	no	84.3	80.6	77.4	79.8	85.1	69.2	66.5	77.6
ZO-SPSA	score	per request	no	82.1	79.4	75.0	76.3	80.7	64.8	62.9	74.5
UltraBreak	none	none	yes	86.7	84.1	81.5	82.4	87.2	73.8	71.6	81.0
Reference Attack	none	none	limited	90.1	88.4	86.2	87.5	91.0	79.3	77.6	85.7
GAMBIT	output	per request	no	88.8	87.1	84.9	86.0	89.4	77.8	75.9	84.3
PolyJailbreak	output	per request	no	91.6	90.3	88.7	89.5	92.1	82.4	80.7	87.9
UniCal	label/none	once/none	yes	90.4	89.7	87.8	88.9	91.2	80.1	77.9	86.6

To avoid searching from scratch on the target, we construct the initial candidate pool entirely from surrogate models. We draw B_0 candidate offsets by sampling a support size uniformly from $\{1, \ldots, s\}$, choosing that support uniformly without replacement, and drawing each active coefficient from a zero-mean Gaussian truncated to $[-a_{\max}, a_{\max}]$. On a surrogate-only proposal split $\mathcal{D}_{\text{prop}}$ disjoint from target calibration and test requests, candidates are ranked by $\widetilde{J}_{\text{sur}}(\alpha) = -K^{-1} \sum_k \mathbb{E}_{r \sim \mathcal{D}_{\text{prop}}} m_k(\text{Compose}(r, c_\alpha)) - \eta \|\alpha\|_1$; the top B form $\mathcal{A}_0$. The values $B_0, B, a_{\max}, \eta$ are fixed by surrogate validation. Let $\mathcal{A}_t$ denote the active proposal set at elimination round t, with $\mathcal{A}_0$ as the initialization. Every proposal in $\mathcal{A}_t$ is evaluated on the same calibration minibatch, ranked by empirical Equation 8, and the lower half is discarded. Let m_t and r_t denote the minibatch size and repetition count in round t, respectively. The total charged target-query budget is $Q_{\text{cal}} = \sum_t |\mathcal{A}_t| m_t r_t = 192$. The optimizer receives only $\ell(o, r)$; response text and continuous evaluator scores remain unavailable. For zero-feedback external transfer, the carrier from the same surrogate-selected proposal is fixed before target access and used on both external models, with no target-side re-ranking or successive elimination.

After calibration, $\hat{\alpha}$ and the corresponding pair $c_{\hat{\alpha}}$ are frozen. For every unseen request, the target receives $F^\star(\text{Compose}(r, c_{\hat{\alpha}}))$, with no request-specific optimization, carrier update, or additional calibration. If calibration costs Q_{cal} target calls and deployment requires Q_{dep} calls per request, the amortized cost over N requests is

$$Q_{\text{amort}}(N) = \frac{Q_{\text{cal}}}{N} + Q_{\text{dep}}. \tag{9}$$

For a per-request baseline with expected cost $Q_b > Q_{\text{dep}}$, the break-even point is $N^\star = \frac{Q_{\text{cal}}}{Q_b - Q_{\text{dep}}}$. We report Q_{cal}, Q_{dep}, and $N^\star$ separately under this accounting rule.

5 Experiments

5.1 Setup and Evaluation Protocol

Data. We use MM-SafetyBench (Liu et al., 2025), VLSU (Palaskar et al., 2026), USB (Zheng et al., 2026), and the VLJailbreakBench released with IDEATOR (Wang et al., 2025a). Requests and source images are partitioned by semantic seed across coordinate learning, surrogate proposal generation, target calibration, and testing. Category and language coverage are tracked separately using surrogate-unseen holdouts, where "unseen" denotes exclusion from coordinate learning and surrogate proposal generation rather than target calibration; calibration and final-test requests remain disjoint by semantic seed. Benign controls come from benchmark-provided safe or borderline examples. Coordinate-map and generator training use surrogate data only.

Models and Baselines. The coordinate pool contains $K = 5$ open surrogates spanning the LLaVA, Qwen2-VL, InternVL2, and MiniCPM-V families (Liu et al., 2023; Wang et al., 2024a; Chen et al., 2024). Representation-level LOMO is used only for coordinate validation in Figure 5 and Appendix Table 10. For the first five targets in Table 1, we use attack-level LOMO. The two evaluated LLaVA checkpoint configurations are denoted LLaVA-A and LLaVA-B. Molmo and Phi-Vision are external unseen architectures excluded from coordinate fitting, generator/proposal construction, and hyperparameter selection. A carrier prepared on the full five-surrogate pool is fixed before access

Table 2: **Closed-target strict success.** API columns average two dated evaluations. Identities are in Appendix Table 6. limited reuse is request-dependent.

Method	Feedback	Adaptation	Reuse	API-A MM	API-A USB	API-B MM	API-B USB	API-C MM	API-C USB	Mean
No attack	n.a.	none	n.a.	6.8	5.2	4.3	3.9	7.5	6.2	5.7
Random search	label	per request	no	34.5	29.8	31.2	26.7	37.4	33.1	32.1
FigStep	none	none	yes	49.6	44.3	46.1	40.8	52.7	48.5	47.0
Con Instruction	none	none	yes	58.2	52.5	54.7	49.6	61.4	56.8	55.5
MML	none	none	limited	87.6	84.2	82.5	79.8	90.1	86.4	85.1
JailBound	none	none	limited	74.8	69.3	72.1	66.4	77.5	71.8	72.0
IDEATOR	output	per request	no	72.5	66.8	68.9	62.7	75.4	70.1	69.4
ZO-SPSA	score	per request	no	67.4	60.3	64.1	57.6	70.8	65.2	64.2
UltraBreak	none	none	yes	76.9	70.5	74.0	67.8	79.2	73.6	73.7
Reference Attack	none	none	limited	91.3	88.7	89.5	86.9	93.1	90.6	90.0
GAMBIT	output	per request	no	90.4	86.8	91.2	87.4	92.6	89.1	89.6
PolyJailbreak	output	per request	no	94.8	92.1	93.5	90.7	95.3	93.0	93.2
UniCal	label	once	yes	83.7	79.9	81.4	76.8	86.1	82.0	81.7

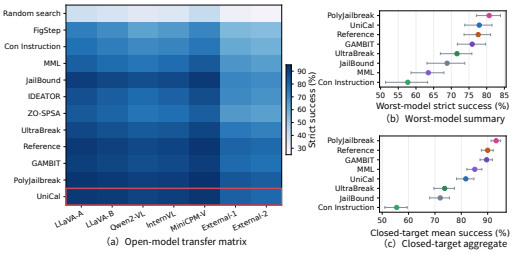


Figure 3: **Representative-baseline transfer.** For UniCal, the first five targets use attack-level LOMO with one-time decision-label calibration, while Molmo/Phi-Vision use zero-feedback transfer with a fixed surrogate-selected carrier; the seven-target worst-model result is descriptive. Error bars show 95% bootstrap CIs. Model-label aliases follow the main text.

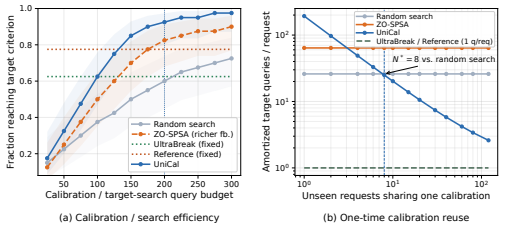


Figure 4: **Calibration success versus query budget and amortized target-query cost under repeated reuse.** In (a), the common criterion is $\widehat{S} \geq \gamma$, where $\widehat{S} = |\mathcal{C}|^{-1} \sum_{r \in \mathcal{C}} \psi(\ell(o_r, r))$ is the mean decision-label score and γ is fixed on surrogate validation data. Shaded bands show confidence intervals. The plot label "richer fb." denotes score feedback for ZO-SPSA.

to either external model and used for both evaluations, with no target-side calibration, re-ranking, or feedback. Closed API targets are external to the surrogate pool; all five open surrogates are used for surrogate-side preparation, followed by one-time decision-label calibration on each dated API evaluation. Closed targets are denoted API-A, API-B, and API-C in the main text. Exact model identities, evaluation dates, region, and access settings are reported in Appendix Table 6, while decoding and repetition settings are fixed within each evaluation round. We compare eleven attacks: Random Search, FigStep (Gong et al., 2025), Con Instruction (Geng et al., 2025), MML (Wang et al., 2025b), JailBound (Song et al., 2025), IDEATOR (Wang et al., 2025a), ZO-SPSA (Guan et al., 2026), UltraBreak (Cui et al., 2026), Reference Attack (Wang et al., 2026b), GAMBIT (Hu et al., 2026), and PolyJailbreak (Wang et al., 2026a). Results are reported by feedback and adaptation regime. The primary query-efficiency track compares UniCal with label-only Random Search under the same three-way interface; native-access baselines retain their original feedback and adaptation rules.

Metrics and Statistics. Strict success requires non-refusal, relevance, and substantive policy violation; benign utility is the preservation rate on benign controls. Detection AUC uses a fixed held-out clean-versus-carrier detector (lower is better; Appendix A.2), distinct from the representation-level covariates. The same frozen evaluator E supplies calibration labels and automated final judgments, but calibration exposes only the three-way label; a stratified human audit assesses evaluator agreement. Confidence intervals are bootstrap-based and multiple comparisons use FDR correction. HarmBench-style judging informs evaluator design (Mazeika et al., 2024).

5.2 Transfer, Universality, and One-Time Reuse

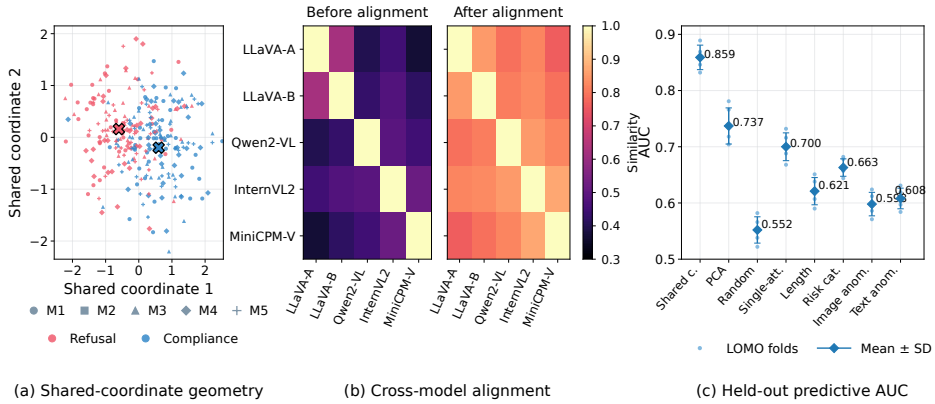
Across five attack-level LOMO open targets, UniCal achieves 89.6% mean strict success after one-time label-only calibration, while the fixed zero-feedback carrier reaches 80.1% on Molmo and 77.9% on Phi-Vision (Table 1 and Figure 3). Since these external models receive no target-side calibration or

Table 3: Calibration ablation.

Variant	Success (%) ↑
Shared z + calib.	**81.7**
Shared z + no calib.	76.6
PCA + calib.	77.4
Single-dir. + calib.	74.9
Random + calib.	65.8

Table 4: Generator-loss ablation.

Variant	Success ↑	Utility ↑	AUC ↓
Full UniCal	81.7	93.5	0.60
w/o $\mathcal{L}_{\text{coord}}$	75.9	92.7	0.63
w/o $\mathcal{L}_{\text{beh}}$	70.8	94.2	0.58
w/o $\mathcal{L}_{\text{vis}}$	79.4	89.8	0.72
w/o $\mathcal{L}_{\text{text}}$	77.6	91.1	0.67
w/o $\mathcal{L}_{\text{util}}$	82.4	85.6	0.64



(a) Shared-coordinate geometry (b) Cross-model alignment (c) Held-out predictive AUC

Figure 5: **Cross-model coordinate validation under representation-level LOMO.** (a) First two dimensions of the learned $d_z = 8$ coordinate; markers denote surrogates and crosses class centroids. (b) Cross-model alignment. (c) Held-out predictive AUC. "Shared c." and "Single-att." denote Shared z and the single-direction control.

re-ranking, the results isolate the transferability of the surrogate-prepared carrier. Representation-level LOMO further yields a mean held-out AUC of 0.859, compared with 0.737 for PCA and 0.700 for the single-direction control, indicating that the learned space preserves cross-model refusal–compliance structure beyond generic low-dimensional variation (Figure 5 and Table 10). On closed targets, one-time calibration achieves 81.7% mean strict success (Table 2), showing that the same calibration mechanism remains effective when target access is restricted to decision labels.

5.3 Analysis

Efficiency and Robustness. UniCal amortizes $Q_{\text{cal}} = 192$ against label-only Random Search ($Q_b = 26$), breaking even after eight requests ($N^\star = 7.68$; Figure 4), with 93.5% benign utility and 0.60 detection AUC (Appendix Table 5). Fixed-carrier transformations average 84.0% strict success (93.5% paraphrase; 73.9% additive noise; Table 12); further trade-offs and evaluator agreement are in Appendix Figure 6 and Table 13.

Ablations. Table 3 shows that both the shared coordinates and calibration matter: removing calibration reduces success from 81.7% to 76.6%, while other alternatives perform worse. Table 4 further shows that L_{beh} and L_{coord} contribute most to attack success, whereas removing L_{util} slightly improves success but reduces benign utility from 93.5% to 85.6%. More ablations are in the Appendix.

6 Conclusion

We presented UniCal, a black-box multimodal jailbreak red-teaming formulation that turns cross-model refusal and compliance representations into a one-time decision-label calibration interface. UniCal Jailbreak achieves 89.6% mean strict success on five attack-level LOMO open targets, 80.1% and 77.9% zero-feedback transfer to Molmo and Phi-Vision, and 81.7% on closed targets with 93.5% benign utility; 192 calibration queries break even after eight reused requests. Future work will extend UniCal to video and tool-augmented agents and study defenses against coordinate-consistent multimodal shifts.

AI USE STATEMENT

Generative AI tools were used for language editing and methodological stress-testing.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Beitao Chen, Xinyu Lyu, shengming yuan, Jingkuan Song, Hengtao Shen, and Lianli Gao. Safeptr: Token-level jailbreak defense in multimodal llms via prune-then-restore mechanism. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 57947–57965. Curran Associates, Inc., 2025.

Renmiao Chen, Yida Lu, Shiyao Cui, Xuan Ouyang, Victor Shea-Jay Huang, Shumin Zhang, Chengwei Pan, Han Qiu, and Minlie Huang. The side effects of being smart: Safety risks in MLLMs' multi-image reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2026.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Intern vl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, 2024.

Kaiyuan Cui, Yige Li, Yutao Wu, Xingjun Ma, Sarah Monazam Erfani, Christopher Leckie, and Hanxun Huang. Toward universal and transferable jailbreak attacks on vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 91–104. IEEE, 2025.

Moreno D'Incà, Nicu Sebe, and Massimiliano Mancini. Harnessing textual refusal directions for multimodal safety, 2026.

Jinhu Fu, Yihang Lou, Qingyi Si, Shudong Zhang, and Sen Su. Diagnosing and repairing unsafe channels in vision-language models via causal discovery and dual-modal safety subspace projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31693–31702, June 2026.

Jiahui Geng, Thy Thy Tran, Preslav Nakov, and Iryna Gurevych. Con Instruction: Universal jailbreaking of multimodal large language models via non-textual modalities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2917–2933, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0.

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 *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8.

Yuxin Gou, Xiaoning Dong, Qin Li, Shishen Gu, Richang Hong, and Wenbo Hu. SURE: Safety understanding and reasoning enhancement for multimodal large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7552–7593, November 2025. ISBN 979-8-89176-332-6.

Jiwei Guan, Haibo Jin, and Haohan Wang. Crafting adversarial inputs for large vision-language models using black-box optimization. In *Findings of the Association for Computational Linguistics: EACL 2026*, March 2026.

Jiahe Guo, Xiangran Guo, Jiaxuan Chen, Weixiang Zhao, Yanyan Zhao, Yutai Hou, Qianchao Wang, Dandan Tu, and Bing Qin. Safety geometry collapse in multimodal llms and adaptive drift correction, 2026.

Mengqi He, Xinyu Tian, Xin Shen, Jinhong Ni, Shu Zou, Zhaoyuan Yang, and Jing Zhang. High-entropy tokens as multimodal failure points in vision-language models. *arXiv preprint arXiv:2512.21815*, 2025.

Mengqi He, Xinyu Tian, Xin Shen, Shu Zou, Jinhong Ni, Zhaoyuan Yang, Weikang Li, Xuesong Li, and Jing Zhang. Break the brake, not the wheel: Untargeted jailbreak via entropy maximization. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=dfX4zpagYF. Expert Certification.

Xiangdong Hu, Yangyang Jiang, Qin Hu, and Xiaojun Jia. GAMBIT: A gamified jailbreak framework for multimodal large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2026.

John Hughes, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Arushi Somani, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, and Mrinank Sharma. Best-of-n jailbreaking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yilei Jiang, Xinyan Gao, Tianshuo Peng, Yingshui Tan, Xiaoyong Zhu, Bo Zheng, and Xiangyu Yue. HiddenDetect: Detecting jailbreak attacks against multimodal large language models via monitoring hidden states. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14880–14893, July 2025. ISBN 979-8-89176-251-0.

Choongwon Kang, Seungjong Sun, Hyunmin Jun, and Jang Hyun Kim. Jailbreaking multimodal large language models using multi-clip video. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *Computer Vision – ECCV 2024*, pp. 386–403, Cham, 2025. Springer Nature Switzerland.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224. PMLR, 21–27 Jul 2024.

Shruti Palaskar, Leon Alexander Gatys, Mona Abdelrahman, Mar Jacobo, Laurence F Lindsey, Rutika Moharir, Gunnar Lund, Yang Xu, Navid Shiee, Jeffrey P. Bigham, Charles Maalouf, and Joseph Yitan Cheng. VLSU: Mapping the limits of joint multimodal understanding for AI safety. In *The Fourteenth International Conference on Learning Representations*, 2026.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21527–21536, 2024. no. 19.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *International Conference on Learning Representations*, volume 2024, pp. 30853–30885, 2024.

Jiaxin Song, Yixu Wang, Jie Li, Xuan Tong, Rui Yu, Yan Teng, Xingjun Ma, and Yingchun Wang. JailBound: Jailbreaking Internal Safety Boundaries of Vision-Language Models. In *Advances in Neural Information Processing Systems*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Ke-Yang Chen, Xue-Jing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Meng-Fei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *ArXiv*, abs/2409.12191, 2024a.

Ruofan Wang, Xingjun Ma, Hanxu Zhou, Chuanjun Ji, Guangnan Ye, and Yu-Gang Jiang. White-box multimodal jailbreaks against large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400706868.

Ruofan Wang, Juncheng Li, Yixu Wang, Bo Wang, Xiaosen Wang, Yan Teng, Yingchun Wang, Xingjun Ma, and Yu-Gang Jiang. Ideator: Jailbreaking and benchmarking large vision-language models using themselves. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8875–8884, 2025a.

Xinkai Wang, Beibei Li, Zerui Shao, Ao Liu, Guangquan Xu, and Shouling Ji. Polyjailbreak: Cross-modal jailbreaking attacks on closed-box multimodal llms. *IEEE Transactions on Dependable and Secure Computing*, 23(5):10789–10806, 2026a.

Yu Wang, Xiaofei Zhou, Yichen Wang, Geyuan Zhang, and Tianxing He. Jailbreak large vision-language models through multi-modal linkage. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1466–1494, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Yulong Wang, Yifei Fu, and Jiayi Gao. Reference attack: A new cross-modal jailbreaking attack against multimodal large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2026b.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: concept cones and representational independence. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Wenzhuo Xu, Zhipeng Wei, Zonghao Ying, Deyue Zhang, Dongdong Yang, Xiangzheng Zhang, and Quanchen Zou. Dmn: A compositional framework for jailbreaking multimodal llms with multi-image inputs, 2026.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qi-An Chen, Hua Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone. *ArXiv*, abs/2408.01800, 2024.

Ziyi Yin, Yuanpu Cao, Han Liu, Ting Wang, Jinghui Chen, and Fenglong Ma. Towards robust multimodal large language models against jailbreak attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22847–22856, June 2026.

Xinyi Zeng, Xue Yang, Jingyuan Zhang, Huanqian Yan, Xiang Chen, Kaiwen Wei, Hankun Kang, and Yu Tian. SafeSteer: A decoding-level defense mechanism for multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, July 2026.

Shenyi Zhang, Keyan Guo, Zihao Wang, Xuebin Li, Lingchen Zhao, Hongxin Hu, Chao Shen, and Qian Wang. MMAligner: Safeguarding multimodal large language models through representation calibration, 2026.

Xinyu Zhang, Daolang Huang, Samuel Kaski, and Julien Martinelli. PABBO: Preferential amortized black-box optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Shiji Zhao, Ranjie Duan, Fengxiang Wang, Chi Chen, Caixin Kang, Shouwei Ruan, Jialing Tao, YueFeng Chen, Hui Xue, and Xingxing Wei. Jailbreaking multimodal large language models via shuffle inconsistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2045–2054, October 2025.

Baolin Zheng, Guanlin Chen, Hongqiong Zhong, Qingyang Teng, Yingshui Tan, Zhendong Liu, Weixun Wang, Jiaheng Liu, Jian Yang, Wenbo Su, Xiaoyong Zhu, Bo Zheng, Kaifu Zhang, Huiyun Jing, and Jincheng Wei. USB: A comprehensive and unified safety evaluation benchmark for multimodal large language models, 2026.

Han Zhu, Juntao Dai, Jiaming Ji, Haoran Li, Chengkun Cai, Pengcheng Wen, Chi-Min Chan, Boyuan Chen, Yaodong Yang, Sirui Han, and Yike Guo. SafeMT: Multi-turn safety for multimodal language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026.