TARGETED TRANSFER ATTACKS ON MLLMS: WHAT A SPECTRAL CONSTRAINT BUYS

Anonymous authors
Paper under double-blind review

ABSTRACT

Targeted adversarial images crafted on open surrogate encoders can transfer to multimodal large language models (MLLMs) that the attacker never queries. The strongest such attacks spend the full ℓ_∞ budget, which leaves open how much perturbation transfer requires and what spending less buys. We study this question through spectral constraints on the perturbation. We insert three rules into one surrogate-only attack skeleton: hard rank truncation, uniform singular-value shrinkage and a gradient-informed threshold. We evaluate them in two regimes. In a high-transfer regime we read held-out encoder success, perceptual distortion and transfer to six generative MLLMs. In a stealth regime at a smaller step we read detectability under a residual-based detector calibrated without any constrained attack. The results separate the skeleton from the rule. Across five attack objectives, the skeleton with no rule reproduces the success gain of the complete configuration over its released baseline, so the rules we ran do not buy success. What a rule shifts is the trade-off at comparable success. In the stealth regime the soft rules lower LPIPS by about one third and reduce detector flagging from near saturation to roughly 0.12–0.14, and the gradient-informed rule still reaches four open-weight MLLMs with every paired difference from the full-budget baseline covering 0. Controls show that neither scaling a perturbation down to the rule's amplitude nor lowering its rank alone reproduces this, and recalibrating the detector on reduced-amplitude copies of the attack it already knows recovers most of the detection, so the evasion depends on detector coverage. Detection and generative transfer are read on different victims. Spectral constraints therefore do not make transfer attacks more successful by themselves. They make transferable perturbations more efficient and less detectable under the studied detector coverage. Code: `https://anonymous.4open.science/r/TASS-0C8E`.

1 INTRODUCTION

A transferable targeted attack on a multimodal large language model (MLLM) crafts an image perturbation on public surrogate models so that a victim the attacker never queries names an attacker-chosen object (Zhao et al., 2023; Li et al., 2025; Hu et al., 2026). It must reach the target on that victim, keep the image change within what the setting tolerates, and pass any input screening. The strongest recipes optimise a free ℓ_∞-bounded tensor against an ensemble of CLIP surrogates (Radford et al., 2021) under heavy augmentation (Xie et al., 2019; Wang et al., 2021; Hu et al., 2026). Their only constraint on the image is the budget $\|\delta\|_\infty \leq \epsilon$, which does not directly control perceptual distortion or spectral allocation, and says nothing about detection.

A success rate does not say how sparing the attack had to be. These recipes spend the whole budget, and an attack that spends less changes the image less and leaves less of the residual signal a steganalyser reads. Such attacks do succeed, so we ask how sparingly one can, and what spending less buys.

Spectral constraints are one way of spending less. Shrinking the singular values of the update concentrates the perturbation into fewer directions and lowers its amplitude with them, and the literature that constrains perturbations by low rank, low frequency or sparsity reports both success and imperceptibility (Guo et al., 2020; Esmaeili et al., 2021; Luo et al., 2022). We insert three such rules into one fixed attack skeleton (Figure 1): hard truncation (HSS), uniform soft thresholding (USS) and a

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.

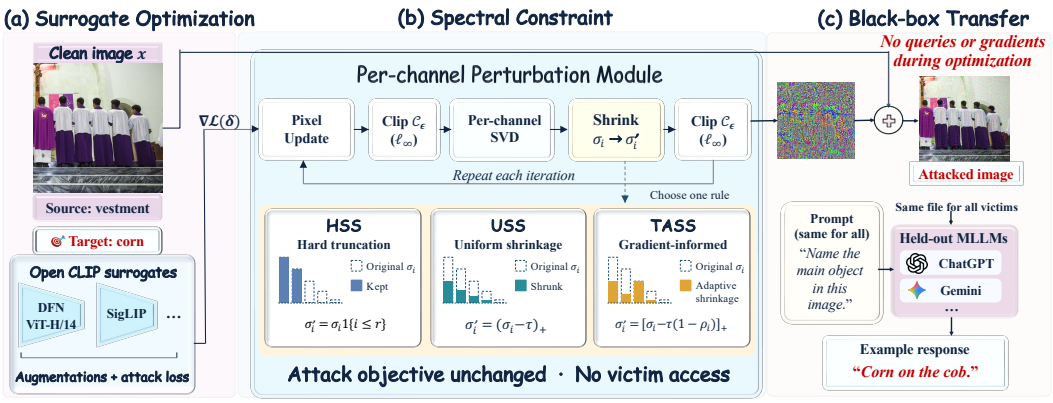

Figure 1: The task and the module: surrogates supply the only gradients, the per-iteration pass clips and then shrinks the singular values, and the attacked image goes as one file to a held-out victim under a neutral prompt.

gradient-informed threshold (TASS). They are the vehicle of the measurement and not a method we propose. On the same images we read three axes at once: target success on two held-out encoders, perceptual distortion, and detectability under a detector calibrated without any constrained arm.

This literature usually compares two complete configurations, which moves the parameterisation, the step size, the energy and the residual statistics a detector reads all at once. A configuration that performs better does not show that the structure contributed, and a low detection rate under one detector does not show that the perturbation is hard to detect under another training coverage or input processing. Each result below is therefore paired with a control that changes one thing: the same skeleton with the rule switched off, a perturbation scaled to the rule's own amplitude, a rule that lowers the rank but not the amplitude, an energy-constrained baseline, and a detector whose training coverage varies at fixed size.

Two terms recur. A paired difference over images is *resolved* when its 95% bootstrap interval excludes 0; an interval that crosses 0 shows neither a difference nor equivalence. A perturbation is *compacted* when its realised energy sits in fewer directions, that is at a lower realised effective rank, in whatever basis. Compaction is a statement about directions, not about amplitude.

The findings are conditional and we state them as such. With the direct-pixel Adam skeleton held fixed, the rule costs a resolved 1.11 [0.39, 1.82] points at step 2.5, and at step 0.5 the two soft settings used for detection cost 1.9 to 3.4 points while lowering LPIPS (Zhang et al., 2018) by about a third. On all five objectives, the skeleton alone with no rule reproduces the gain that the complete spectral configurations show over each objective's released configuration (§5.3). Because that change bundles the parameterisation with the step, the gain belongs to the skeleton as a whole rather than to direct-pixel updates alone. None of the rules we ran buys success, and the six generative victims show the same configuration-level gain, with no consistent ordering among the rules (§5.2).

On the detection axis the rule does carry the effect. At comparable held-out-encoder success on the same images, a zero-day steganalyser (Kodovsky et al., 2011; Fridrich & Kodovsky, 2012) flags the unconstrained perturbation on 0.98, the same skeleton with no rule on 0.94, and the soft rules on 0.12–0.14, and that ordering recurs on ImageNetV2 and in a second, modified pipeline (§5.4, §5.6). A perturbation scaled to the rule's own amplitude is flagged 0.26 to 0.80 more often and loses 17 to 67 points of success, so spending less is not by itself enough. Lowering the realised rank by 36 at fixed amplitude leaves the rate unchanged, so compacting is not enough either. Neither rank nor amplitude alone accounts for the detection results, and we do not identify which property of the realised spectrum does.

The advantage is a property of the evaluation as much as of the attack. Calibration that includes the compacted family, or merely the known family rescaled to lower energy, recovers detection, and bicubic resizing of the input changes what the same detector reads (§5.6). We contribute:

- **Problem.** We ask what spectral constraints change in targeted transfer attacks under a fixed ℓ_∞ budget, reading success, perceptual distortion and detectability together, and separating the constraint's effect from the optimisation configuration and the detector's calibration.
- **Method.** We build one surrogate-only attack skeleton that hosts hard truncation and uniform or gradient-informed spectral shrinkage, with a no-rule baseline, amplitude-matched constructions and rank-reduction controls that test the competing explanations.
- **Experiments.** We run five attack objectives on 500 image–target pairs, read on two held-out encoders, six generative MLLMs and two supervised classifiers, with a 500-image replication on a second dataset (ImageNetV2), a second attack pipeline, two budgets and a seventeen-cell rule grid. The rules we ran buy distortion and detectability, not success, and only inside one skeleton under the stated detector calibration and preprocessing.

2 Related Work

Targeted transfer attacks. Momentum (Dong et al., 2018), input diversity and transformations (Xie et al., 2019; Dong et al., 2019; Lin et al., 2020; Wang et al., 2021), and gradient variance reduction (Wang & He, 2021) all improve transfer, and attacks on CLIP-based and multimodal systems (Radford et al., 2021; Zhao et al., 2023; Hu et al., 2026) build on this recipe. Omni-Attack itself attacks open-ended VQA and OCR with its own target construction; we reuse only its optimisation recipe as our baseline. Targeted instance attacks change the surrogate loss (Zhao et al., 2021) or its features (Byun et al., 2023; Wei et al., 2023; Liang et al., 2025; Jia et al., 2025), or they train a generator (Naseer et al., 2021; Fang et al., 2024; Zhang et al., 2025), and further work continues along these axes (Yang et al., 2024; Zeng et al., 2025; Huang et al., 2025; Xu et al., 2025; Li et al., 2026; Su et al., 2026). Four of these objectives run as losses inside our pipeline (App. F).

Structure and energy constraints. Low-frequency perturbations transfer well in the untargeted setting (Sharma et al., 2019; Guo et al., 2020), and frequency constraints reduce distortion (Luo et al., 2022). Structured parameterisations range from low-rank-plus-sparse subspaces (Esmaeili et al., 2021) and sparse structure (Croce & Hein, 2019) to a factorised per-channel low-rank PGD (Savostianova et al., 2024) and low-rank query subspaces (Liu et al., 2026). These constrain different objects: a matrix spectrum is not a frequency band, a factorisation fixes the rank before optimisation where a post-update rule shapes a free tensor, and a query subspace serves a query attack. Singular-value thresholding is the proximal map of the nuclear norm (Cai et al., 2010) and is exactly our uniform rule; its weighted form is weighted nuclear-norm minimisation (Gu et al., 2014), in the reweighting tradition of Candes et al. (2008). Energy is instead constrained by a joint L_2 penalty (Carlini & Wagner, 2017b) or a perceptual threat model (Zhao et al., 2020; Laidlaw et al., 2021); we report RMS and LPIPS (Zhang et al., 2018) descriptively and claim no imperceptibility. The factorised UV attack is not compared under a matched protocol here (App. A).

Detection and its evaluation. Our detector pairs a spatial rich-model residual feature set (Fridrich & Kodovsky, 2012) with an ensemble of Fisher linear discriminants (Kodovsky et al., 2011), chosen because it reads the image alone and is trained without gradients from any victim. For adversarial example detectors it is established that a detector's verdict depends on the attack families it was trained on, and that adaptive attacks and distribution shift defeat detectors that looked strong (Carlini & Wagner, 2017a). We do not add a new detector; what we add is evidence of how the detection verdict on structured perturbations moves when training coverage and input processing change.

3 Problem Setup

Task. Given a clean image x whose main object is g and an attacker-chosen target class $t \neq g$, the attacker seeks δ with $\|\delta\|_\infty \leq \epsilon$ such that a held-out generative victim identifies the main object as t when it is asked a neutral question about $x + \delta$ with no target label in the prompt. The victim is never queried and contributes no gradient; the attack optimises on an ensemble of open CLIP encoders, following Hu et al. (2026), and the final image must lie in the valid pixel range. This is an input attack on object identification rather than a text jailbreak, and the setting is surrogate-only and instance-wise throughout.

Three quantities, kept apart. *Effectiveness* is whether an unqueried victim outputs the target. Generative success is targeted ASR, the fraction of pairs on which the free-form response identifies the main object as t. Encoder success, used for the larger-scale analysis, is 1000-way target top-1 on two held-out CLIP encoders. *Image change* is the per-pixel RMS of the released perturbation on the 0 to 255 scale and LPIPS (Zhang et al., 2018). A lower value is not a claim of imperceptibility. *Detection* is whether a detector flags the image under a stated training coverage, threshold and input preprocessing: a joint property of the attack distribution and the detection system, reported only with those conditions. We compute the spectral-entropy effective rank $\exp(-\sum_i p_i \log p_i)$ with $p_i = \sigma_i / \sum_j \sigma_j$ per colour channel on the released perturbation and average it (App. B). It is an auxiliary descriptor, not a goal. The two held-out encoders never enter the optimisation but informed early design choices, so we read every encoder number as mildly optimistic; App. C splits them into development and test.

Surrogate objective. The augmentations are Gaussian noise, random crop and pad, differentiable JPEG and patch dropout. At each step, for each surrogate m and a random augmentation $\mathcal{A}$ drawn from these, the loss pulls the augmented embedding toward target references and away from ground-truth references:

$$\mathcal{L}(\delta) = \mathbb{E}_{m,\mathcal{A}}\Big[\ell_{\text{align}}\big(f_m(\mathcal{A}(x+\delta)),\, t,\, g\big)\Big], \tag{1}$$

which we minimise. The clips of §4 and Algorithm A1 enforce the bound $\|\delta\|_\infty \le \epsilon$ on the final perturbation.

Models and data. We optimise on five CLIP surrogates (DFN ViT-H/14 at 378 and 224, SigLIP SO400M at 384 and 224, SigLIP-L/16 at 384). We evaluate on two held-out CLIP victims (ViT-bigG/14, CLIPA ViT-H/14-336), six generative victims (§5.1), and, for a larger domain gap, supervised ImageNet classifiers (ResNet-50, ConvNeXt-Base). Images are the NIPS-2017 development set (Kurakin et al., 2018) at 299×299 with its fixed target annotation, and every arm, budget and seed attacks the same image–target pairs. A replication uses 500 matched-frequency images of ImageNetV2 (Recht et al., 2019), which shares the label space (§5.4, App. G). Attacks run 1500 steps at ϵ=16/255 unless stated otherwise. We compute every metric on the evaluated perturbation $\delta_{\text{eval}} = x_{\text{adv}} - x$, where x_{adv} is the released 8-bit image. As in the base attack's release, iterates after step 200 are averaged at rate 0.01 and clipped. We store δ dense, so no rank figure implies a memory saving.

4 Method

The module acts on the perturbation after each Adam update of the base attack, and keeps that attack's loss, surrogates and augmentations. Let δ^k be the perturbation at the start of iteration $k = 0, \dots, K-1$, let $\widehat{\mathcal{L}}_k$ be one sampled draw of Eq. 1, and let a^k be the Adam state. One iteration is then

$$D^k = \nabla_\delta \widehat{\mathcal{L}}_k(\delta^k), \qquad (\widetilde{\delta}^{k+1}, a^{k+1}) = \text{Adam}(\delta^k, D^k, a^k), \qquad \delta^{k+1} = \mathcal{Q}_{\text{rule}}(\widetilde{\delta}^{k+1}), \tag{2}$$

where $\mathcal{Q}_{\text{rule}}$ is the pass of one rule. Each pass is built with the clip $\mathcal{C}_\epsilon$, which sends every entry to $[-\epsilon, \epsilon]$. The Adam state is kept across the pass, the next gradient is taken at δ^{k+1}, and no rule queries a victim or computes an extra gradient. Inside the attack, pixels are on the 0 to 255 scale, so ϵ=16, τ and the Adam step are all in these units. Each rule has one strength parameter, either r or τ, and it works on every colour channel $Z \in \mathbb{R}^{H \times W}$ of every image separately (Figure 1). It acts through the thin SVD $Z = \sum_{i=1}^{q} \sigma_i\, u_i v_i^\top$, where $\sigma_1 \ge \cdots \ge \sigma_q \ge 0$ and $q = \min(H, W)$, and it replaces the coefficients σ_i and keeps u_i and v_i.

4.1 Three spectral rules

Hard spectral shrinkage (HSS) keeps the r largest components of each channel and removes the rest. Its pass alternates this truncation with the clip three times:

$$\text{HSS}_r(Z) = \sum_{i=1}^{r} \sigma_i\, u_i v_i^\top, \qquad \mathcal{Q}_{\text{hss}} = (\mathcal{C}_\epsilon \circ \text{HSS}_r)^{\circ 3}, \tag{3}$$

where the first truncation acts on the unclipped update, and each one decomposes its current input afresh. Every truncation has rank at most r, but the clip after it can raise the rank, so r bounds neither the clipped nor the released perturbation (App. A). *Uniform spectral shrinkage* (USS) subtracts one threshold $\tau \geq 0$ from every singular value, which is standard singular-value thresholding (Cai et al., 2010). Its pass clips once and shrinks once:

$$\mathrm{USS}_\tau(Z) = \sum_{i=1}^{q} (\sigma_i - \tau)_+ \, u_i v_i^\top, \qquad \mathcal{Q}_{\mathrm{uss}} = \mathrm{USS}_\tau \circ \mathcal{C}_\epsilon, \tag{4}$$

where $(a)_+ = \max(a, 0)$. Shrinking singular values does not in general keep every entry in $[-\epsilon, \epsilon]$, so the iterate can leave the budget set until the next step clips it. At $\tau{=}0$ the pass reduces to $\mathcal{C}_\epsilon$, which is the no-rule control. *Target-aware spectral shrinkage* (TASS) lets the surrogate gradient, and never a victim, lower the threshold of the components it loads on. Let $G^{k+1} = \gamma\, G^k + (1{-}\gamma)\, D^k$ with $G^0 = 0$, and let G be the matching channel of G^{k+1}. Then

$$\mathrm{TASS}_\tau(Z; G) = \sum_{i=1}^{q} \big(\sigma_i - \tau(1 - \rho_i)\big)_+ \, u_i v_i^\top, \qquad \rho_i = \frac{s_i}{\max_j s_j + \epsilon_{\mathrm{num}}}, \tag{5}$$

where $s_i = |u_i^\top G\, v_i|$, $\gamma{=}0.9$ and $\epsilon_{\mathrm{num}}{=}10^{-12}$, and $\max_j$ runs over that channel's components, so $0 \leq \rho_i < 1$. The pass is $\mathrm{TASS}_\tau(\cdot\,; G^{k+1}) \circ \mathcal{C}_\epsilon$, and $G = 0$ recovers USS. The loading drops the sign of the projection and averages past gradients, so it is a weighting heuristic, not the loss change of removing a component (App. A). Because $\tau(1 - \rho_i) \leq \tau$, the two soft rules at one τ are not at one operating point (App. C). The realised structure of a released perturbation depends jointly on the rule, its strength, the update that replenishes the spectrum between passes, and the export averaging and clipping. No nominal setting describes it on its own.

5 Experiments

5.1 Protocol and metrics

Objectives, victims, and configurations. Five attack objectives run in the one pipeline: Omni-Attack (Hu et al., 2026), FTM (Liang et al., 2025), FOA-Attack (Jia et al., 2025), M-Attack (Li et al., 2025) and MF-ii (Zhao et al., 2023). Encoder results use the two held-out CLIP encoders of §3 on 500 image–target pairs. Six generative victims receive the final images of 200 pairs, frozen before any query: GPT-5.6-Sol and Gemini-3.8-flash through their providers' APIs, and Qwen2.5-VL-7B-Instruct (Bai et al., 2025), Qwen3.8-27B, DeepSeek-V4.1-flash (Xu et al., 2026) and GLM-5.3-flash (Zeng et al., 2026) as open weights (protocol in App. E). The baseline configuration of each objective runs its loss with the $\epsilon \tanh$ re-parameterisation of the released Omni-Attack code and Adam at step 0.05. The spectral configurations run in two regimes, and every result is reported in its own regime. The *high-transfer* regime takes direct-pixel updates at step 2.5, with $r{=}128$ for HSS and $\tau{=}1$ for USS and TASS, and success and transfer to the generative victims are read there. The *stealth* regime takes step 0.5, at $\tau{=}1$ for TASS, $\tau{=}0.5$ for USS and $r{=}128$ for HSS, and detection is read there. The TASS threshold was selected on a development victim and checked on a disjoint test victim, with the step fixed beforehand (App. B), and the USS threshold is the sweep point nearest TASS in realised RMS (App. C). Every arm shares the pairs and the settings of §3 at $\epsilon{=}16/255$, fixed rather than per-rule optima.

Metrics, statistics, and the detector. Encoder success is aggregated over the two victims, each pair contributing its two outcomes as one sampling unit, and both rates are in percent. Generative labels come from a rule-based matcher that an independent blinded pass agrees with on 89.4% of 18,000 responses (App. E). We pair differences over pairs and form 95% bootstrap intervals from 10^4 resamples over images rather than over attack seeds, unless stated otherwise. The detector of §2 is calibrated out of fold on a clean pool and the unconstrained attack only, the *zero-day* setting: no constrained arm ever enters training, thresholds come from a separate clean development pool, and the realised false-positive rate from a further one (App. D). Success, distortion and detection are always reported on the same scored images, and arms are compared by the paired difference over them.

Table 1: Targeted transfer to six generative MLLMs on 200 image–target pairs at ϵ=16/255. ASR counts responses naming the main object as the target and bold is the best point estimate per group at the fixed settings.

		Closed-source ASR (%) ↑		Open-weight ASR (%) ↑				
Base attack	Variant	GPT-5.6	Gemini-3.8	Qwen2.5	Qwen3.8	DeepSeek	GLM-5.3	LPIPS ↓
Omni-Attack (Hu et al., 2026)	Baseline	26.5	32.0	36.5	43.5	40.0	35.0	0.346
	+ HSS	29.0	**39.5**	37.0	**47.5**	45.0	40.0	**0.235**
	+ USS	27.5	37.0	42.0	45.0	41.5	39.0	0.238
	+ TASS	**31.0**	35.0	**42.5**	46.5	**47.5**	**44.0**	0.248
FTM (Liang et al., 2025)	Baseline	26.5	36.5	36.5	41.0	36.5	30.5	0.352
	+ HSS	31.5	**39.5**	**41.0**	42.0	**47.5**	40.0	**0.233**
	+ USS	30.5	36.5	38.0	43.0	45.5	36.5	0.236
	+ TASS	**33.5**	38.0	40.0	**45.0**	44.5	**40.5**	0.245
FOA-Attack (Jia et al., 2025)	Baseline	11.0	17.0	20.0	20.5	19.0	21.5	0.334
	+ HSS	14.5	21.5	24.5	25.5	23.5	23.5	**0.224**
	+ USS	18.0	22.5	**27.5**	27.0	**31.5**	**27.5**	0.228
	+ TASS	**18.5**	**24.0**	25.5	**27.5**	27.0	23.5	0.233
M-Attack (Li et al., 2025)	Baseline	11.5	17.0	21.0	22.5	21.0	17.5	0.338
	+ HSS	17.0	**22.5**	24.0	25.5	24.0	25.0	**0.225**
	+ USS	**18.5**	22.0	22.5	25.5	**24.5**	24.5	0.227
	+ TASS	16.5	20.5	**29.5**	**27.5**	23.5	**28.0**	0.238
MF-ii (Zhao et al., 2023)	Baseline	11.5	18.0	20.0	20.5	19.0	19.0	0.335
	+ HSS	15.0	21.5	23.0	23.0	27.0	23.5	**0.224**
	+ USS	**18.0**	18.5	26.5	**28.0**	27.0	22.0	0.228
	+ TASS	16.5	**22.0**	**27.0**	26.5	**29.0**	**26.5**	0.239

5.2 Transfer to generative victims

Table 1 sends the final images of the five objectives, in the high-transfer regime, to the six generative victims. Clean images never name the target (App. E). The objectives fall into two bands: Omni-Attack and FTM reach 26.5 to 43.5 from their baseline configurations, against 11.0 to 22.5 for FOA-Attack, M-Attack and MF-ii. The victims separate as widely. GPT-5.6-Sol is the hardest column throughout, between 11.0 and 33.5, and the four open-weight victims sit about seven points above the two closed-source ones. Distortion moves the same way on every objective and by almost the same amount, LPIPS falling from 0.334–0.352 at the baseline to 0.224–0.235 under HSS, between 32 and 34% on all five, with HSS lowest in every group while rarely highest on ASR. Relative to the baseline configuration of the same objective, over the 30 objective–victim cells the ASR point estimate is higher in 30, 29 and 30 cells for HSS, USS and TASS, and the interval excludes 0 in 6, 11 and 16 of them. Between the two soft rules, one cell of 30 excludes 0. The resolved cells are marginal intervals, and an independent re-labelling preserves the sign of the baseline-to-rule difference in 54 of 60 re-checked cells (App. E). The table therefore shows a configuration-level gain that the three rules share, with no consistent ordering among them. At a second attack seed the Omni-Attack baseline-versus-TASS pair reverses sign on Gemini-3.8-flash (-3.0 $[-10.5, +4.5]$). The stealth-regime TASS arm of the Omni-Attack objective, sent to the four open-weight victims on the same pairs, prompt and rule, is not distinguished from the baseline configuration on any of the four, every paired difference covering 0, and on two of them sits a resolved -10.5 and -9.0 points below the high-transfer arm, so the regimes differ in transfer and we do not claim they are equivalent (App. E).

5.3 Effects under fixed optimisation

Does the success gain of a complete spectral configuration belong to the rule? Table 2 holds the direct-pixel Adam skeleton fixed at step 2.5 and switches the module off or to each rule on every objective, which supplies the no-rule control that question needs. The skeleton alone reproduces the gain. Without any rule it is worth 4.4 to 8.4 points over the released configuration on all five objectives, every interval excluding 0. The baseline is not under-tuned, since its step of 0.05 is the

Table 2: Does the success gain belong to the rule? Target top-1 (%) on the two held-out encoders and LPIPS, 500 pairs at ϵ=16/255: the released configuration of each objective, the direct-pixel Adam skeleton at step 2.5 with no rule, and that skeleton with each rule, differing in the rule alone. Bold is a column's best within an objective. Table A1 adds RMS, detection and paired intervals.

	Omni-Attack		FTM		FOA-Attack		M-Attack		MF-ii	
Configuration	Top-1	LPIPS	Top-1	LPIPS	Top-1	LPIPS	Top-1	LPIPS	Top-1	LPIPS
Released (ϵ tanh, step 0.05)	85.4	0.340	84.1	0.344	42.6	0.328	42.9	0.331	41.8	0.329
Skeleton, no rule (τ=0)	**90.7**	0.252	**88.5**	0.249	48.0	0.245	**49.0**	0.245	**50.2**	0.245
+ HSS (r=128)	89.4	**0.229**	88.1	**0.226**	46.8	**0.217**	46.5	**0.218**	46.4	**0.218**
+ USS (τ=1)	89.6	0.231	87.7	0.229	**48.6**	0.220	**49.0**	0.220	49.2	0.221
+ TASS (τ=1)	**90.7**	0.241	86.7	0.238	46.7	0.225	48.8	0.231	48.3	0.231

Table 3: Is a compacted perturbation detected less often at comparable success? Zero-day protocol of §5.1, Omni-Attack objective; every column is averaged over that panel's scored images and detection rates are fractions with cluster-bootstrap intervals. Joint is the fraction of them reaching the target on both held-out encoders and not flagged at 5%. (a) NIPS-2017. (b) ImageNetV2. Bold marks a column's best within a panel, ties included; it marks no overall winner, and the energy control leads on both distortion columns in both panels.

Configuration	Top-1 (%) ↑	LPIPS ↓	RMS ↓	Eff. rank	TPR@5% ↓	Joint ↑
(a) NIPS-2017, Omni-Attack objective						
Omni-Attack (ϵ tanh, step 0.05)	92.0	0.355	12.80	221.3	0.98 [0.94, 1.00]	0.00 [0.00, 0.00]
None (τ=0, step 0.5)	92.0	0.298	10.46	215.5	0.94 [0.86, 1.00]	0.04 [0.00, 0.10]
HSS (r=128, step 0.5)	92.0	0.275	9.97	139.2	0.84 [0.74, 0.94]	0.10 [0.02, 0.18]
USS (τ=0.5, step 0.5)	**93.0**	0.195	7.36	133.3	**0.12** [**0.04**, **0.22**]	**0.80** [**0.68**, **0.90**]
TASS (τ=1, step 0.5)	89.0	0.191	7.29	124.5	0.14 [0.06, 0.24]	0.76 [0.64, 0.86]
Omni-Attack+L_2 (ball 7.33, energy)	91.0	**0.178**	**6.55**	207.8	0.32 [0.20, 0.44]	0.60 [0.46, 0.74]
(b) ImageNetV2, Omni-Attack objective, every image scored						
Omni-Attack (ϵ tanh, step 0.05)	88.1	0.341	12.70	221.3	1.00 [0.99, 1.00]	0.00 [0.00, 0.01]
None (τ=0, step 0.5)	**89.8**	0.283	10.40	215.6	0.95 [0.94, 0.97]	0.04 [0.02, 0.06]
HSS (r=128, step 0.5)	89.6	0.260	9.91	141.3	0.35 [0.30, 0.39]	0.60 [0.55, 0.64]
USS (τ=0.5, step 0.5)	89.7	0.179	7.27	133.5	**0.07** [**0.05**, **0.09**]	**0.81** [**0.77**, **0.84**]
TASS (τ=1, step 0.5)	88.0	0.176	7.22	125.2	**0.07** [**0.05**, **0.09**]	0.80 [0.76, 0.83]
Omni-Attack+L_2 (ball 7.33, energy)	87.2	**0.163**	**6.53**	208.0	0.09 [0.07, 0.12]	0.75 [0.71, 0.79]

best of the four we swept and the direct-pixel step is the implementation's default (App. A), and the no-rule arm is not buying success with distortion, since on every objective it sits 25 to 28% below the released configuration in LPIPS and in RMS alike. Adding a rule does not add to it. Fourteen of the fifteen objective-by-rule cells lie at or below their own no-rule row, and all three intervals that exclude 0 exclude it downwards (HSS on M-Attack and on MF-ii, TASS on FTM). Most cells are individually too small to resolve, and an interval covering 0 is not evidence of equivalence, but pooled over the fifteen and paired over the same 500 pairs the rule costs a resolved 1.11 [0.39, 1.82] points. What it buys instead is distortion, 4 to 11% below its own no-rule row in both LPIPS and RMS, every LPIPS reduction resolved. Step 2.5 is the rules' most favourable reading, since at step 0.5, where §5.4 reads detection, the three rules sit 2.2, 8.5 and 3.4 points below the no-rule arm on Omni-Attack, and the two soft settings the detection tables use sit 1.9 and 3.4 below it (App. C).

The advantage is therefore the skeleton's, on all five objectives rather than inferred from one. Two limits: we claim only that none of the rules we ran buys success, not that no spectral rule can, and because the skeleton change bundles the parameterisation with the step, the gain belongs to the skeleton as a whole and not to direct-pixel updates alone. The rule grid, the τ sweeps, the target-aware variants and a second attack seed are in App. C.

5.4 Detection at comparable success

Table 3a reads success, distortion and detection on one set of scored images. Its four step-0.5 rows differ in the rule alone, and the unconstrained attack, the no-rule control and hard truncation all reach top-1 92 on the cohort. Figure 2 shows what the five configurations write onto one image of that cohort at equal success and equal peak amplitude, so what separates them is distortion and spectrum. The L_2 control carries *less* realised energy than either soft rule at a far higher rank, so it is a control

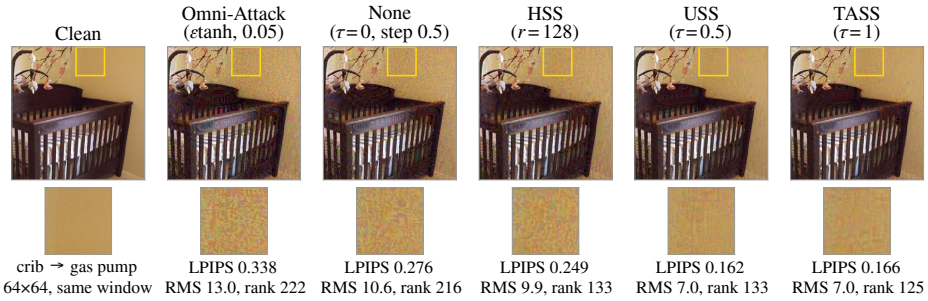

Figure 2: What the attack writes onto one image, on the Omni-Attack objective, for the five configurations. All five reach the target on both held-out encoders at $\max|\delta|=16$, so the columns differ in distortion and in spectrum, not in success or in peak amplitude. The pair and the window are drawn blind (App. H).

Table 4: Does the rule do anything a smaller perturbation would not? Each block fixes a per-pixel RMS and varies only how a perturbation of that amplitude was reached. Δ is paired against each block's first row. USS is the one rule with a complete record at both amplitudes; TASS agrees, HSS reaches neither, and a third band at RMS ≈ 10.4 catches every construction (App. C).

Construction	Top-1 (%) ↑	Δ top-1 (pp) [95%]	RMS	Eff. rank	TPR@5% ↓
Matched at RMS ≈ 7.3					
USS (τ=0.5), as run	93.0	–	7.36	133.3	0.12
None (τ=0), scaled	76.0	−17.0 [−27.0, −8.0]	7.21	215.7	0.80
Omni-Attack baseline, scaled	54.0	−39.0 [−51.0, −27.0]	7.08	221.6	0.92
Matched at RMS ≈ 5.7					
USS (τ=1), as run	86.0	–	5.69	101.5	0.04
None (τ=0), scaled	41.0	−45.0 [−57.0, −33.0]	5.75	215.6	0.30
Omni-Attack baseline, scaled	19.0	−67.0 [−78.0, −56.0]	5.73	221.1	0.76

for energy rather than a twin matched on it. It keeps the released $\epsilon \tanh$ parameterisation, so it cannot isolate the spectrum, and no comparison in this paper separates the rule from that parameterisation.

At 5% false positives the detector flags the unconstrained attack and the no-rule control alike. Paired, the soft rules sit -0.82 $[-0.92, -0.70]$ and -0.80 $[-0.90, -0.68]$ below the no-rule control and are not distinguished from each other, $+0.02$ $[-0.04, +0.10]$, while hard truncation buys almost nothing, -0.10 $[-0.18, -0.02]$ at r=128. Against the energy control the compacted arm, TASS, is -0.18 $[-0.30, -0.06]$ at 5%, and against the unconstrained attack -0.84 $[-0.94, -0.74]$. But that control carries less energy than either soft rule and is flagged more often than both, so what the rule buys inside the skeleton the $\epsilon \tanh$ parameterisation does not buy outside it at the same energy. Panel (b), which scores all 500 attacked images against panel (a)'s 50, repeats the pattern on ImageNetV2 with the compacted arm minus the energy control at -0.03 $[-0.04, -0.01]$, so the sign of the gap is stable across collections and its size is not. Under a fixed skeleton, then, the rule produces the low rate.

5.5 Amplitude or spectrum?

The arms of Table 3 differ in amplitude as well as in rule, so Table 4 fixes the amplitude and varies only how a perturbation of that amplitude was reached. At RMS $\approx$ 7.3 the same pipeline with no rule, scaled to USS's amplitude, is flagged $+0.68$ $[+0.54, +0.80]$ more often, and the released attack scaled to it $+0.80$ $[+0.68, +0.90]$ more. At RMS $\approx$ 5.7 this repeats at $+0.26$ and $+0.72$, and both bands cost 17 to 67 points of success. The rule did not produce the success of a spectral configuration (§5.3), but it produces this, and turning the amplitude down is no substitute on either axis.

Neither rank nor amplitude alone accounts for the table. The scaled rows reach the rule's amplitude at rank 216–222 against its 102–133 and are flagged far more often, so at a fixed amplitude the re-

alised spectrum matters. But rank is not a dial either. In the rule grid of App. C, HSS at $r{=}192$ takes the rank from 216 to 180 at unchanged amplitude and moves detection by $+0.00$ $[+0.00, +0.00]$, and the grid's lowest-rank hard-truncation cell, HSS at $r{=}96$, is flagged $+0.32$ $[+0.20, +0.46]$ *more* than a USS cell 50 rank points above it that carries less amplitude. Nor is the singular basis shown to be necessary: a DCT-compacted arm is flagged about as often as the singular-basis arm, which resolves no difference and establishes no equivalence, while a random basis at comparable compaction and energy is flagged three times as often, so compaction as such is not what evades (App. D). What the rule contributes is some property of the realised spectrum, of which the effective rank is only a summary, and not merely a lower amplitude. Which property it is, we do not identify. The rule grid and the matched-amplitude controls exist on one objective only.

5.6 Scope and replication

What the detector had seen. The low rate is a property of the evaluation as much as of the attack. At the same splits and the same budget of attacked training images, changing only which family the detector trains on, a calibration set that includes the compacted family raises detection of it by a paired $+0.68$ $[+0.56, +0.80]$, and one in which half the attacked images are the *known* attack at a quarter of its amplitude, and nothing else is new, gains $+0.62$ $[+0.48, +0.76]$ at almost no cost on the original attack, from 0.26 to 0.88 at a nominal 10% false-positive rate (Table A4). What the compacted arm evades is a detector calibrated on one amplitude of one attack. Whether the recovery reaches a family the defender can neither have seen nor generate is not established.

Preprocessing. The native-224 MPCAttack images are flagged on 0.98, and the same images upsampled to 299 by bicubic interpolation and scored by the same frozen detector on 0.00 (App. D). We have not measured the attenuation of the high-pass residual that the natural reading assumes, so what we establish is narrower: a frozen detector's verdict changes completely under an input transformation that leaves the attack intact, and every arm is scored at the resolution it was produced at, on a detector not re-fitted on resized images.

Replication. The ordering of Table 3 recurs on ImageNetV2 (panel b) and in a second attack pipeline, MPCAttack (Li et al., 2026) run in its own code and surrogate pool at 224 px with its sign step replaced by Adam, where the module lowers a rate of 1.00 to 0.10 $[0.02, 0.20]$ at a success difference over the 500 pairs of $+1.4$ $[-1.7, +4.4]$ (Table A3, App. F). It is not a reproduction of MPCAttack, and the two replications are not a complete cross-design. In the native sign-step update the module neither raises success nor lowers detection, and at step 2.5 TASS lowers detection of the four ported objectives only to 0.72–0.82 (Table A1, App. F). A second budget and two supervised victims, where the CLIP parity stops at that domain gap, are in App. G. Success and distortion are read on the encoders and the generative victims alike, and agree, while detection is read on the encoders alone, so no single setting is measured as both effective and unflagged on the same victims.

6 Conclusion

We asked how sparingly a targeted transfer attack on an MLLM can succeed and what spending less buys. The answer separates the skeleton from the rule. On five objectives the skeleton alone, with no rule, reproduces the success gain of a complete spectral configuration, and six generative MLLMs show the same configuration-level gain with no consistent ordering among the rules. What a rule buys at comparable success is distortion and detectability: about a third less LPIPS, and a zero-day detection rate of 0.12–0.14 against 0.94 with no rule, an ordering that recurs on ImageNetV2 and in a second pipeline (§5.2–5.4, §5.6). That rate is not explained by a smaller perturbation or a lower rank alone (§5.5), and it is as much a property of the evaluation as of the attack, since recalibrating on reduced-amplitude copies of the known attack raises detection of the compacted arm from 0.26 to 0.88 at a nominal 10% false-positive rate (§5.6). All of this holds for the rules we ran, in one skeleton at two fixed settings, on one image collection with a 500-image replication, one detector family, one budget, two held-out encoders that informed early design and one attack seed for most arms.

Reproducibility Statement

All code, including attack implementations and evaluation scripts, is available in the repository linked in the abstract, with experimental settings and analysis protocols detailed in the main text and appendices. We also provide per-image results and query records to support verification, in the Supplementary Material.

AI use statement

We used generative AI tools in this work for three things: polishing the text of the manuscript, debugging the experimental, analysis and auditing code, and cross-checking derivations. We did not use them to generate, select or alter any experimental result. The authors reviewed all AI-assisted content. Every numeric claim in the paper is in addition machine-checked against the stored experimental artifacts by an automated audit. We take responsibility for the final content of this work, including the text, the claims and the artifacts produced with the aid of these tools.

Ethics statement

This work studies targeted adversarial transfer to multimodal large language models to inform robustness evaluation, while acknowledging that reduced perturbation distortion and detectability may facilitate misuse. Adversarial images were submitted to two commercial MLLM APIs (GPT-5.6-Sol and Gemini-3.8-flash) as single-image queries under a neutral prompt. No deployed application or other user was targeted, and every query is recorded in the released records. Our experiments use established image benchmarks without targeting user data or downstream applications, and we report the conditions and limitations of our findings to support responsible interpretation and defense research.

References

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.

Junyoung Byun, Myung-Joon Kwon, Seungju Cho, Yoonji Kim, and Changick Kim. Introducing competition to boost the transferability of targeted adversarial examples through clean feature mixup. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24648–24657, 2023.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier analysis and applications*, 14(5):877–905, 2008.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security (AISec)*, pp. 3–14, 2017a.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 39–57, 2017b.

Francesco Croce and Matthias Hein. Sparse and imperceivable adversarial attacks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4723–4731, 2019.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9185–9193, 2018.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4307–4316, 2019.

Ashkan Esmaeili, Marzieh Edraki, Nazanin Rahnavard, Mubarak Shah, and Ajmal Mian. Ls-dat: Low-rank and sparse decomposition for decision-based adversarial attack. *arXiv preprint arXiv:2103.10787*, 2021.

Hao Fang, Jiawei Kong, Bin Chen, Tao Dai, Hao Wu, and Shu-Tao Xia. Clip-guided generative networks for transferable targeted adversarial attacks. In *European Conference on Computer Vision (ECCV)*, pp. 1–19. Springer, 2024.

Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on information Forensics and Security*, 7(3):868–882, 2012.

Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2862–2869, 2014.

Chuan Guo, Jared S Frank, and Kilian Q Weinberger. Low frequency adversarial perturbation. In *Uncertainty in Artificial Intelligence Conference (UAI)*, pp. 1127–1137, 2020.

Kai Hu, Weichen Yu, Li Zhang, Alexander Robey, Andy Zou, Haoqi Hu, Chengming Xu, and Matt Fredrikson. Omni-attack: Adversarial attacks on open-ended vqa in black-box multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42341–42351, 2026.

Hanxun Huang, Sarah Monazam Erfani, Yige Li, Xingjun Ma, and James Bailey. X-transfer attacks: Towards super transferable adversarial attacks on CLIP. In *International Conference on Machine Learning (ICML)*, 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 57993–58028, 2025.

Jan Kodovsky, Jessica Fridrich, and Vojtěch Holub. Ensemble classifiers for steganalysis of digital media. *IEEE Transactions on information forensics and security*, 7(2):432–444, 2011.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. In *The NIPS'17 Competition: Building Intelligent Systems*, pp. 195–231. Springer, 2018.

Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *International Conference on Learning Representations (ICLR)*, 2021.

Yuanbo Li, Tianyang Xu, Cong Hu, Tao Zhou, Xiao-Jun Wu, and Josef Kittler. Multi-paradigm collaborative adversarial attack against multi-modal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 45018–45051, 2025.

Kaisheng Liang, Xuelong Dai, Yanjie Li, Dong Wang, and Bin Xiao. Improving transferable targeted attacks with feature tuning mixup. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2020.

Han Liu, Shanghao Shi, Yevgeniy Vorobeychik, Chongjie Zhang, and Ning Zhang. Low rank adaptation for adversarial perturbation. *arXiv preprint arXiv:2604.27487*, 2026.

Cheng Luo, Qinliang Lin, Weicheng Xie, Bizhu Wu, Jinheng Xie, and Linlin Shen. Frequency-driven imperceptible adversarial attack on semantic similarity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15294–15303, 2022.

Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. On generating transferable targeted perturbations. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7688–7697, 2021.

Sen Nie, Jie Zhang, Jianxin Yan, Shiguang Shan, and Xilin Chen. V-attack: Targeting disentangled value features for controllable adversarial attacks on lvlms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, pp. 5389–5400. PMLR, 2019.

Dayana Savostianova, Emanuele Zangrando, and Francesco Tudisco. Low-rank adversarial pgd attack. *arXiv preprint arXiv:2410.12607*, 2024.

Yash Sharma, Gavin Weiguang Ding, and Marcus A. Brubaker. On the effectiveness of low frequency perturbations. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

Tongrui Su, Qingbin Li, Shengyu Zhu, Wei Chen, and Xueqi Cheng. Rapa: Enhancing transferable targeted attacks via random parameter pruning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6538–6548, 2026.

Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1924–1933, 2021.

Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16138–16147, 2021.

Zhipeng Wei, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Enhancing the self-universality for transferable targeted attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12281–12290, 2023.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2725–2734, 2019.

Anyi Xu, B Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, CC Yu, et al. Deepseek-v4. 1-flash: Pushing the limits of kv cache compression. *arXiv preprint arXiv:2609.19969*, 2026.

Binyan Xu, Xilin Dai, Di Tang, and Kehuan Zhang. One surrogate to fool them all: Universal, transferable, and targeted adversarial attacks with clip. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 3087–3101, 2025.

Hunmin Yang, Jongoh Jeong, and Kuk-Jin Yoon. Facl-attack: Frequency-aware contrastive learning for transferable adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6494–6502, 2024.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Hui Zeng, Sanshuai Cui, Biwei Chen, and Anjie Peng. Two heads are better than one: Averaging along fine-tuning to improve targeted transferability. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19900–19909, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zhengyu Zhao, Zhuoran Liu, and Martha Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1036–1045, 2020.

Zhengyu Zhao, Zhuoran Liu, and Martha Larson. On success and simplicity: A second look at transferable targeted attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

APPENDIX CONTENTS

Each section below holds one subject, so that a claim in the main text can be checked without reading the rest. Section letters and page numbers are cross-referenced, and every table, figure and algorithm in the appendix is numbered A1 onwards.

	Section	What it holds	Page
A	Method and implementation	The optimiser and operator details, the shrinkage schematic, Algorithm A1, why the rule is a post-update pass rather than a factorisation.	14
B	Evaluation metrics and statistics	How success, distortion and the paired intervals are computed.	16
C	Controls under fixed optimisation	The five-objective rule table, the complete rule grid, the module against no rule, the τ sweeps, the third amplitude band, the target-aware variants, a second attack seed, the energy control, and an HSS pilot across budgets.	16
D	Detection protocol and controls	The detector and its out-of-fold discipline, the two replications, the basis and shared-cap controls, and what the defender had to have seen.	19
E	Generative-victim evaluation	The query protocol frozen before any victim was seen, the label review, the second-seed pair, the rule-off generative control, the native comparison, and two recorded response pairs.	21
F	External baselines and scope	The ported losses, the native sign-step pipelines, and the attacks we do not compare against.	25
G	Generalisation across budgets, datasets and victims	A second budget, the ImageNetV2 replication, and supervised CNN victims.	25
H	Additional visual examples	Three pairs under all five objectives with identical crops, and how they were selected.	26

A METHOD AND IMPLEMENTATION

Success rates and their paired differences are printed in percent throughout, as in the main text. Figure A1 contrasts the three rules on one illustrative spectrum. In Figure 1 the example response at the right is the recorded one for that frozen pair under the TASS configuration.

Optimisation. The baseline configuration of every attack uses the $\epsilon \tanh$ re-parameterisation of the released Omni-Attack code at Adam step 0.05, which was the strongest of the swept values, since the sweep over 0.025, 0.05, 0.1 and 0.5 gives top-1 87, 93, 91 and 51 on 50 images. The no-rule control, HSS, USS and TASS instead optimise the pixel tensor directly with Adam. Every $n{=}500$ run in this paper uses batch size 25. At that batch size the implementation's default step is $0.1B$, which gives 2.5, while the step-0.5 arms set 0.5 explicitly, so the two branches differ in step size only. The step size sets how much spectrum is replenished between passes. At $\epsilon{=}16$ and $\tau{=}1$, for instance, TASS ends at a lower effective rank at step 0.5 than at step 2.5 (Table A2). Omni-Attack$+L_2$ is the baseline configuration with a per-image projection onto an L_2 ball after each step. Its ball radii (7.33, 9.0, 11.0) are in the same per-pixel RMS units as the distortion columns of Table 3, and a run's measured RMS sits at or below its radius, at 6.54 and 10.42 for the 7.33 and 11.0 balls, averaged over the run. The SVDs are batched over the three colour channels. On the same cluster and at the same batch size, one 50-image arm takes about one GPU hour for the baseline and for the soft rules, and HSS takes 4 to 5 times that. These are wall-clock ratios rather than an efficiency claim.

Operators, budget, and bounds. The HSS pass of Eq. 3 starts from $Z^{(0)} = \widetilde{\delta}^{k+1}$ and iterates $Z^{(j+1)} = \mathcal{C}_\epsilon(\mathrm{HSS}_r(Z^{(j)}))$ for $j = 0, 1, 2$, each step decomposing its own current input. The bound $\mathrm{rank} \le r$ holds after each truncation and before the clip that follows it; it does not hold for the output, for the iterate average or for the saved image, and the alternation is not an exact projection onto the intersection of the rank set and the budget set. Smaller singular values do not imply smaller entries, so iterates produced by a soft rule can leave the budget set between two clips; this follows from the operator rather than from floating-point error, and we do not report how often it occurs.

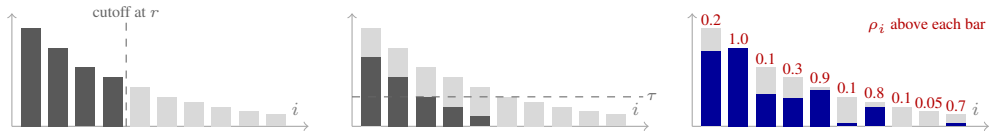

Figure A1: Spectral shrinkage rules on an illustrative spectrum. HSS applies a hard cutoff, USS a uniform soft threshold, and TASS adjusts the shrinkage of each direction by its normalised gradient loading ρ_i (red). Grey and coloured bars show the input and retained values; the values are illustrative, not measured.

Algorithm A1 Spectral-constraint module inside the attack loop (as implemented)

Require: clean image x, target t, budget ϵ, surrogates m=1..5, rule $\in$ {none, hss, uss, tass} with parameter r or τ (none runs uss at τ=0), steps K=1500; $\delta^0 \leftarrow 0$ (Adam parameter), EMA gradient $G^0 \leftarrow 0$, γ=0.9, ϵ_{num}=10^{-12}

for $k = 0, \ldots, K-1$ **do**
  $D^k \leftarrow \nabla_\delta \widehat{\mathcal{L}}_k(\delta^k)$ {one EOT draw: surrogates × augmentations of $x + \delta^k$, each augmented image clipped to the valid range; the base attack's loss}
  $(\widetilde{\delta}^{k+1}, a^{k+1}) \leftarrow \mathrm{Adam}(\delta^k, D^k, a^k)$ {the Adam state is never reset by the spectral pass}
  $G^{k+1} \leftarrow \gamma\, G^k + (1-\gamma)\, D^k$ {read by tass only}
  if rule = hss **then**
    $Z \leftarrow \widetilde{\delta}^{k+1}$ {no clip before the first truncation}
    repeat three times: $Z_c \leftarrow \mathrm{HSS}_r(Z_c)$ for each channel c (a fresh SVD of the current Z_c), then $Z \leftarrow \mathrm{clip}(Z, -\epsilon, \epsilon)$
  else
    $Z \leftarrow \mathrm{clip}(\widetilde{\delta}^{k+1}, -\epsilon, \epsilon)$
    for each colour channel c **do**
      $U\Sigma V^\top \leftarrow \mathrm{SVD}(Z_c)$
      $\tau_i \leftarrow \tau$ (uss, none) or $\tau_i \leftarrow \tau(1-\rho_i)$ with ρ_i from G_c^{k+1} by Eq. 5 (tass)
      $\sigma'_i \leftarrow \max(\sigma_i - \tau_i, 0)$; $Z_c \leftarrow U\,\mathrm{diag}(\sigma')\,V^\top$ {no clip after this pass}
    end for
  end if
  $\delta^{k+1} \leftarrow Z$
  $\bar{\delta} \leftarrow \delta^{k+1}$ if k=200; $\bar{\delta} \leftarrow \bar{\delta} + 0.01\,(\delta^{k+1} - \bar{\delta})$ if k>200 {iterate average of the base attack's release protocol}
end for
$x_{\text{adv}} \leftarrow \mathrm{clip}_{[0,255]}\big(x + \mathrm{clip}(\bar{\delta}, -\epsilon, \epsilon)\big)$, rounded to integers and saved as PNG
every metric uses $\delta_{\text{eval}} = x_{\text{adv}} - x$ read back from the PNG

Because the clean pixels and ϵ are integers, the rounding of Algorithm A1 keeps the released image within the budget. The no-rule control runs the USS code at τ=0 (Eq. 4), which equals $\mathcal{C}_\epsilon$ up to the floating-point error of one SVD and one reconstruction. Before any clip, a TASS pass keeps every coefficient at least as large as USS does at the same τ; this orders neither the final distortion, the final rank nor the success of two runs. Its loading uses the absolute projection of an averaged gradient, whereas the first-order loss change from removing component i, $-\sigma_i\, u_i^\top \nabla\mathcal{L}(Z)\, v_i$, depends on the sign, on σ_i and on the gradient at Z, which is why we read the loading as a weighting heuristic. With frozen, order-preserving weights the soft rule would be the proximal map of a weighted nuclear norm (Gu et al., 2014), and with $w_i \equiv 1$ standard singular-value thresholding (Cai et al., 2010); as run it refreshes the weights every step, clips before it shrinks, and inverts the order of a mean 12.9% of adjacent singular-value pairs along the recorded τ=1 trajectory, so that reading does not apply to the algorithm and no result of the paper depends on it.

Why a post-update rule rather than a factorisation. A factorisation fixes the rank before optimisation and therefore searches a smaller set, whereas a post-update rule lets the attack optimise a free tensor and shapes what the attack produces. We take the second route so that the arms of §5.3

can differ in the rule alone. We ran no matched comparison against the per-channel UV parameterisation of Savostianova et al. (2024), and we claim none.

B Evaluation metrics and statistics

Encoder success. For each victim encoder we build class prototypes from 50 reference embeddings per ImageNet class. Target top-1 requires the target class to be the most similar of the 1000 classes, whereas the target-versus-ground-truth win (tvg) requires only that the top-10 mean similarity to the target references exceed the same average computed over the ground-truth references. For supervised classifiers the win rate compares the target logit with the ground-truth logit. Clean images give near-zero rates under every one of these criteria.

Distortion. RMS is the per-pixel root mean square of the released perturbation δ_{eval} on the 0 to 255 scale, as the main-text equation defines it, and LPIPS uses the AlexNet backbone. Both quantities are computed between the clean image and the evaluated adversarial image. Effective rank is the spectral-entropy effective rank of each colour channel of δ_{eval}, and we average it over the channels and over the first 100 released perturbations of the run.

Paired differences and launches. The per-comparison intervals for every cell of Table A1 and Table 1 are in the released per-image records. Every paired difference is computed within one launch and one table; two seeds exist for Omni-Attack, for the three Omni-Attack+L_2 balls and for the step-2.5 TASS points, and the seed-1 launch is in App. C. The comparisons of this paper are exploratory: no difference was pre-registered, and the detector's splits and thresholds were frozen while the arms scored against them were not. No multiplicity correction is applied, and a cell that resolves on its own is read as exactly that.

Selection of the stealth regime. The τ of the stealth regime is selected on a development victim and checked on a disjoint test victim. With CLIPA-H as development and ViT-bigG as test, we sweep $\tau \in \{1, 2\}$ and take the setting with the lowest LPIPS that stays within a 2-point non-inferiority margin of the baseline. That rule picks τ=1, which reaches top-1 91.0 against the baseline's 93.0, whereas τ=2 is ineligible at 82.0. On the test victim the frozen choice then gives top-1 91.0 against 92.0 at LPIPS 0.188 against 0.347 (n=100).

C Controls under fixed optimisation

The grid. Table A2 holds the rule grid on this objective at the two steps, so the rows Table A1 and Table 3a print are visibly cells of a sweep rather than chosen points. Over the seventeen cells, realised amplitude and realised rank correlate with the detection rate at Spearman +0.90 and +0.75, and partialling each out leaves +0.89 for amplitude against +0.74 for rank: amplitude is the better single predictor, rank is far from irrelevant, and neither ordering is what the matched-amplitude controls of Table 4 measure. The matched pairs are sharper. HSS at r=192 on the step-0.5 branch lowers the rank by 36 at an amplitude difference of 0.02 and changes nothing; USS at τ=0.25 carries a rank 50 points *above* HSS at r=96 and is flagged half as often, at the lower amplitude of the two.

The third amplitude band. Table 4 omits the band at RMS $\approx$ 10.4, where the construction makes no difference because everything is caught. Its three rows, on the 50 scored ids, are the no-rule arm as run (92.0 top-1, rank 215.5, flagged 0.94), HSS at r=192 as run (92.0, rank 179.8, 0.94; paired against the first, +0.00 [+0.00, +0.00]), and the released attack scaled to that amplitude (90.0, rank 221.4, 0.96). Lowering the realised rank by 36 while leaving the amplitude alone is what the second of these does, and it changes nothing. TASS sits with USS in the two bands Table 4 prints: at RMS 7.29 it reaches top-1 89.0 at rank 124.5 and is flagged 0.14, +0.02 [−0.04, +0.10] from USS; at RMS 5.66, τ=2 reaches 88.0 and is flagged 0.04, +0.00 [+0.00, +0.00] from it; that arm's perturbations were not retained, so its rank is not recomputable.

Parameterisation and shrinkage rule. The USS rows at τ=0.5 on the step-0.5 branch and at τ=1 on the step-2.5 branch are the sweep points whose realised RMS is nearest that of TASS on

Table A1: What a spectral rule changes, on every attack objective. One direct-pixel Adam pipeline at step 2.5, 500 pairs at ϵ=16/255, seed 0; within an objective the rule rows differ in the rule alone, and every Δ, the released row's included, is paired against that block's None row. That row is not a rule-free control: it differs in the parameterisation and the step as well, so its Δ is the skeleton's gain with the sign turned round. Settings are fixed, not per-rule optima (§5.1); the step-0.5 branch and the grid are in Table A2, and Table 2 prints the top-1 and LPIPS columns. Top-1, LPIPS and RMS are over the 500 pairs; TPR is over the 50 ids the frozen detector scores, at the step-2.5 settings of this table. Bold marks a column's best within an objective, not an overall winner: a rule leading on distortion usually trails on success.

Spectral rule	Top-1 (%) ↑	Δ top-1 (pp) [95%]	LPIPS ↓	RMS ↓	TPR@5% ↓
Omni-Attack (Hu et al., 2026)					
released (ϵ tanh, step 0.05)	85.4	−5.3 [−7.5, −3.2]	0.340	12.75	0.98
None (τ=0), *reference*	**90.7**	–	0.252	9.44	0.94
HSS (r=128)	89.4	−1.3 [−2.9, +0.3]	**0.229**	8.63	**0.64**
USS (τ=1)	89.6	−1.1 [−2.7, +0.5]	0.231	**8.59**	0.78
TASS (τ=1)	**90.7**	+0.0 [−1.4, +1.4]	0.241	8.89	0.82
FTM (Liang et al., 2025)					
released (ϵ tanh, step 0.05)	84.1	−4.4 [−6.7, −2.1]	0.344	12.85	0.96
None (τ=0), *reference*	**88.5**	–	0.249	9.37	0.92
HSS (r=128)	88.1	−0.4 [−2.2, +1.5]	**0.226**	8.57	**0.68**
USS (τ=1)	87.7	−0.8 [−2.5, +0.9]	0.229	**8.52**	0.76
TASS (τ=1)	86.7	−1.8 [−3.5, −0.1]	0.238	8.83	0.82
FOA-Attack (Jia et al., 2025)					
released (ϵ tanh, step 0.05)	42.6	−5.4 [−8.3, −2.7]	0.328	12.53	0.96
None (τ=0), *reference*	48.0	–	0.245	9.33	0.94
HSS (r=128)	46.8	−1.2 [−3.7, +1.2]	**0.217**	8.39	0.66
USS (τ=1)	**48.6**	+0.6 [−1.5, +2.7]	0.220	**8.34**	**0.62**
TASS (τ=1)	46.7	−1.3 [−3.5, +0.9]	0.225	8.47	0.76
M-Attack (Li et al., 2025)					
released (ϵ tanh, step 0.05)	42.9	−6.1 [−8.7, −3.6]	0.331	12.48	0.96
None (τ=0), *reference*	**49.0**	–	0.245	9.28	0.92
HSS (r=128)	46.5	−2.5 [−4.7, −0.3]	**0.218**	8.38	**0.56**
USS (τ=1)	**49.0**	+0.0 [−2.0, +2.0]	0.220	**8.29**	0.60
TASS (τ=1)	48.8	−0.2 [−2.4, +1.9]	0.231	8.64	0.72
MF-ii (Zhao et al., 2023)					
released (ϵ tanh, step 0.05)	41.8	−8.4 [−11.6, −5.3]	0.329	12.54	0.96
None (τ=0), *reference*	**50.2**	–	0.245	9.33	0.94
HSS (r=128)	46.4	−3.8 [−6.4, −1.3]	**0.218**	8.40	0.60
USS (τ=1)	49.2	−1.0 [−3.3, +1.3]	0.221	**8.33**	**0.56**
TASS (τ=1)	48.3	−1.9 [−4.1, +0.3]	0.231	8.69	0.76

the same branch (7.40 against 7.35, and 8.59 against 8.89). The sweep was run on the test cohort, so the choice is post hoc, and the effective ranks are matched only approximately. At step 0.5, HSS at r=128 sits between the soft rules on success (86.1 against 86.4 and 84.9) at a much higher LPIPS (0.263 against 0.186 and 0.183): at this step the hard cut removes fewer small directions per step than a soft threshold does, so it is the least compressive rule on that branch and the most compressive one on the step-2.5 branch.

Module against no rule. Three things are left out of the Δ columns of Table A1, whose fifteen cells pooled and paired over the same 500 pairs give the rule a cost of 1.11 [0.39, 1.82] points. First, all fifteen Δ LPIPS intervals exclude 0 with no endpoint further than 0.004 from its point estimate, so every distortion reduction there is resolved, and at step 2.5 TASS also lowers per-pixel RMS below the no-rule arm by 0.55 [0.54, 0.57]. Second, the soft rules have to be compared with each other at the nearest matched RMS rather than each with None, because the intervals of Table A1 share a None row and are correlated. At step 0.5 the difference USS minus TASS is +1.5 [−0.3, +3.4] on top-1 and +0.003 [+0.001, +0.004] on LPIPS; at step 2.5 it is −1.1 [−2.7, +0.4] and −0.009 [−0.010, −0.009], while the τ=2 USS point matches TASS's success (+0.1 points, interval crossing 0) at lower LPIPS and RMS. At the point estimates USS at τ=2 is therefore ahead of TASS at τ=1 on both metrics, but no success difference between the soft rules is resolved and no

Table A2: The complete rule grid on the Omni-Attack objective. One direct-pixel Adam pipeline, one seed; within a step size the cells differ in the rule and in nothing else. Top-1, LPIPS and RMS are over the 500 pairs, as in Table A1; effective rank and the two detection rates are over the 50 ids the frozen detector scores, as in Table 3a, which is why a rank here can differ from the same arm's rank in Table 3b by up to about two points. A dagger marks the rows another table prints: in the step-2.5 block the four direct-pixel Omni-Attack rows of Table A1, and in the step-0.5 block the four direct-pixel rows of Table 3a together with the USS row Table 4 anchors its lower block on. HSS at r=96 and at r=192 and USS at τ=0.25 carry no dagger because no table prints them, although §5.5 quotes all three from here.

	500 pairs				50 scored ids	
Spectral rule	Top-1 (%)	LPIPS	RMS	Eff. rank	TPR@5%	TPR@10%
Direct-pixel Adam, step 2.5						
None (τ=0)†	90.7	0.252	9.44	214.5	0.94	0.96
HSS (r=96)	88.7	0.219	8.35	134.8	0.36	0.70
HSS (r=128)†	89.4	0.229	8.63	158.8	0.64	0.86
HSS (r=192)	91.3	0.244	9.16	192.6	0.92	0.94
USS (τ=1)†	89.6	0.231	8.59	193.9	0.78	0.90
USS (τ=2)	90.8	0.211	7.89	176.8	0.26	0.64
TASS (τ=0.5)	91.7	0.246	9.15	203.8	0.90	0.96
TASS (τ=1)†	90.7	0.241	8.89	195.1	0.82	0.92
TASS (τ=2)	92.2	0.230	8.49	181.0	0.44	0.84
Direct-pixel Adam, step 0.5						
None (τ=0)†	88.3	0.284	10.45	215.5	0.94	0.96
HSS (r=96)	83.7	0.249	9.58	115.0	0.66	0.88
HSS (r=128)†	86.1	0.263	9.96	139.2	0.84	0.94
HSS (r=192)	88.3	0.281	10.43	179.8	0.94	0.98
USS (τ=0.25)	88.0	0.233	8.70	165.1	0.34	0.50
USS (τ=0.5)†	86.4	0.186	7.40	133.3	0.12	0.28
USS (τ=1)†	79.8	0.116	5.74	101.5	0.04	0.18
TASS (τ=1)†	84.9	0.183	7.35	124.5	0.14	0.26

equivalence is claimed; under the frozen development/test protocol below, the pooled target-aware family trails USS on MF-ii by -2.8 $[-5.2, -0.4]$. Third, HSS at r=128 is a third point on the same trade-off: within this launch HSS minus TASS is -1.3 $[-3.0, +0.4]$ on success and -0.012 $[-0.013, -0.011]$ on LPIPS, at effective rank 160 against 196, so at these settings the hard rule buys the most compression at a success cost this launch does not resolve. A post-hoc intervention on 30 converged Omni-Attack perturbations found that keeping the 128 highest-loading directions retains more of the attack than keeping random or lowest-loading ones (63.3 against 16.7 and 0.0 top-1), so the loading is predictive of where the attack lives; that does not translate into a gain over uniform shrinkage, and the paper does not rest on it.

Trade-off sweeps on every objective. We swept USS and TASS over $\tau \in \{0.5, 1, 2\}$ on each objective at step 2.5, alongside the HSS point and the baseline configuration; the per-point records are released with the code. On every objective the two soft rules interleave: at the same τ, TASS shrinks less and sits at slightly higher LPIPS (by $+0.005$ at τ=0.5, $+0.005$ to $+0.011$ at τ=1 and $+0.019$ to $+0.021$ at τ=2), and its top-1 difference from USS lies between -2.0 and $+1.5$ points across the 15 comparisons, with no interval excluding 0. Varying the nominal rank of HSS moves it along the same curve while the compression is moderate (r=96, 128 and 192 reach 88.7, 89.4 and 91.3 at LPIPS 0.219, 0.229 and 0.244 on Omni-Attack, at effective ranks 138, 160 and 192), but under strong compression it falls below it: r=64 and r=48 give 84.3 and 81.8 at LPIPS 0.204 and 0.197 (ranks 109 and 92), whereas the soft rules keep 88.1 (USS at τ=4, LPIPS 0.174, rank 151) and 87.9 (TASS at τ=8, LPIPS 0.188, rank 141) at the same or lower LPIPS, and FTM repeats this (79.5 and 83.5 for the hard cuts at LPIPS 0.195 and 0.202, against 84.4 at 0.171 and 85.8 at 0.185 for the soft rules). Cutting directions is therefore costlier than shrinking them once the compression is strong. At the moderate settings several pairs have one rule ahead on both point estimates, but at the same τ no pair combines a resolved success difference with lower distortion, which is the sense in which the rules are points on one curve.

Per-objective selection under a development/test split. Because a fixed τ is not a matched operating point, we also select τ per objective under a protocol that was fixed before any test number was read. The development victim is CLIPA-H and the test victim is ViT-bigG. Among the settings $\tau \in \{0.5, 1, 2, 4, 8\}$ at step 2.5, a setting is admissible if its LPIPS does not exceed that of USS at τ=1 on the same objective by more than 0.005, and for each rule we choose the admissible setting with the highest development-victim top-1. On the test victim the selected TASS minus the selected USS is $+0.4$ $[-1.6, +2.4]$ for Omni-Attack (τ=2 against τ=2), $+2.0$ $[+0.2, +4.0]$ for FTM (2 against 2), -2.0 $[-5.2, +1.2]$ for FOA-Attack (2 against 1), -3.2 $[-6.0, -0.6]$ for M-Attack (2 against 1), and -2.4 $[-5.6, +0.8]$ for MF-ii (2 against 1). Under matched distortion the target-aware rule is therefore ahead on one class-target objective and behind on one image-target objective, and the difference is unresolved on the other three.

Variants of the target-aware rule. To test whether a different form of the target-aware weight, rather than a different τ, separates TASS from USS, we ran four groups of variants at n=500 and step 2.5: (i) weights $(1-\rho_i)$ rescaled to mean one per channel, so that the nominal thresholds average to τ as in USS, which does not thereby equalise the mass removed; (ii) the strong-compression settings $\tau \in \{4, 8\}$ for USS, TASS and the normalised variant; (iii) the discount shapes $(1 - \rho_i)^p$ with $p \in \{0.5, 2, 4\}$, rank and quantile normalisation of the loadings, and a hard target-aware cut that leaves the 30% or 60% most-loaded directions untouched, on Omni-Attack and FTM; (iv) step 5. Comparing every target-aware point with every USS point of the same objective whose LPIPS lies within 0.010 gives 65 paired differences on the pooled victims, of which one excludes zero: Omni-Attack TASS at τ=2 against USS at τ=1, $+2.6$ $[+0.8, +4.4]$ at LPIPS 0.230 against 0.231. The normalised variant matches USS at every τ (on Omni-Attack at τ=1 the difference is $+1.1$ $[-0.5, +2.7]$, and the other objectives stay within one point); at τ=8 USS falls to 79.9 against 87.9 for TASS but at LPIPS 0.115 against 0.188, and at matched LPIPS the strong settings give -1.8 $[-3.6, 0.0]$ on Omni-Attack and -0.6 on FTM; the shape and normalisation variants move along the same curve (p=4 at τ=1 reaches 91.2 at 0.248), and step 5 lowers every rule (89.2 with no rule, 87.9 for TASS). Pooling all target-aware variants into one candidate set under the development/test protocol above selects plain TASS at τ=2 on four objectives and the normalised variant on MF-ii, leaves the first four test-victim differences against USS unchanged at $+0.4$, $+2.0$, -2.0 and -3.2, and moves MF-ii from -2.4 to -2.8 $[-5.2, -0.4]$.

Second attack seed. The Omni-Attack four-variant group was rerun at attack seed 1 on the same 500 pairs. It reaches 86.0 for the baseline, 90.5 for direct pixel with no rule, 90.5 for USS, 90.0 for HSS and 90.8 for TASS, at LPIPS 0.340, 0.253, 0.233, 0.230 and 0.240 in that order. TASS minus the baseline is $+4.8$ $[+2.8, +6.8]$, TASS minus no rule is $+0.3$ $[-1.5, +2.1]$, TASS minus USS is $+0.3$ $[-1.3, +1.9]$ and TASS minus HSS is $+0.8$ $[-0.9, +2.5]$. These reproduce the seed-0 picture, where the same four differences against the same launch are $+5.3$, $+0.0$, $+1.1$ and $+1.3$.

Energy control. An L_2 penalty under Adam has no usable range, so the energy control is a free perturbation re-optimised inside a per-image L_2 ball, at radii 7.33, 9.0 and 11.0. Table 3 prints the 7.33 arm; at radius 11.0 the control reaches top-1 87.7 at LPIPS 0.281, RMS 10.42 and effective rank 219.2, and on a second seed the highest Omni-Attack+L_2 success on any rung is 87.8 against 87.7. Reading it against a rule arm at equal success is not a distortion claim across families, because a lower effective rank is not a proxy for lower distortion; the comparison that holds the amplitude fixed instead is Table 4.

HSS across budgets. In a 50-image pilot under the tvg criterion, with the HSS step chosen on success only, HSS at nominal rank 128 matches the free attack at three budgets (90.0 against 88.0 at ϵ=8, 98.0 against 98.0 at ϵ=16, 100.0 against 100.0 at ϵ=32) at LPIPS 0.119, 0.274 and 0.489 against 0.167, 0.357 and 0.627, which is 22% to 29% lower, with RMS lower by 22% to 27%. The nominal rank is imposed on the intermediate tensor and the final clip raises the realised rank, so these are operating points, not a rank guarantee.

D Detection protocol and controls

Detector and protocol. The detector is the standard strong steganalysis configuration of §2, run under the zero-day protocol of §5.1. An SRM-style residual bank computes about 700 features

from the image alone, with no clean reference, and an ensemble of Fisher linear discriminants on random feature subspaces then classifies those features. Scoring is out of fold, so each attacked image receives exactly one score, from a detector that never saw it and that was never calibrated on a constrained arm. The clean development pool that fixes the thresholds enters no training fold. The cluster bootstrap over attacked image ids uses 5000 resamples. We report AUC and TPR at 5% and 10% false positives. We do not report TPR at 1%, because its threshold would rest on three development images. Pooling scores over repeated random splits would instead let one image enter the bootstrap many times, which narrows the intervals without adding information, and we therefore do not use it.

Replications. Table 3b runs the same protocol on ImageNetV2 with its own frozen splits and five folds; the clean pool holds 900 images disjoint from the 500 attacked ones, and the realised false-positive rates are 0.057 and 0.137 at the nominal 5% and 10%, so the 10% rates for (b) are read at a slightly looser operating point than in (a). At 10% the rows of Table 3 read: in (a), 0.98 for the released configuration and 0.60 for the Omni-Attack+L_2 control, with the four step-0.5 rows in Table A2; in (b), 1.00, 0.99, 0.15, 0.14 and 0.24 for the released configuration, the no-rule arm, USS, TASS and the Omni-Attack+L_2 control. One scored ImageNetV2 arm was left out of panel (b) to align the two panels, TASS at τ=1 and step 2.5: top-1 90.2, LPIPS 0.237, RMS 8.82, effective rank 195.3, flagged on 0.35 [0.31, 0.40] at 5% and 0.86 at 10%, joint rate 0.57 [0.52, 0.61]. It reproduces the step-2.5 block of Table A2: the same rule at the same τ is caught far more often when the larger step leaves the realised rank near 195.

Table A3: The same comparison in a second attack pipeline, which runs the MPCAttack objective and surrogates in their own code with the sign step replaced by direct-pixel Adam at step 0.5, and reads images at their native 224 px. Conventions, scored ids and frozen detector as in Table 3; 500 pairs, 50 scored.

Configuration	Top-1 (%) ↑	LPIPS ↓	RMS	Eff. rank	TPR@5% ↓	Joint ↑
None (τ=0)	45.0	0.384	13.65	170.0	1.00 [1.00, 1.00]	0.00 [0.00, 0.00]
TASS (τ=0.25)	49.0	0.177	8.34	73.8	0.10 [0.02, 0.20]	0.32 [0.20, 0.44]
TASS (τ=1)	39.0	0.075	4.23	60.2	0.08 [0.02, 0.16]	0.20 [0.10, 0.32]

Table A3 scores the MPCAttack arms, produced at their native 224 px, on the main-collection splits. There a rate of 1.00 falls to 0.10 [0.02, 0.20] at a success difference over the 500 pairs of +1.4 [−1.7, +4.4], the table's top-1 column being the 50 scored ids. Over all 500 pairs the realised effective ranks are 170.1 with no rule, 74.3 at τ=0.25 and 60.9 at τ=1, within a point of the scored-cohort values, so this pipeline reaches a far lower rank at a far smaller τ than the main one does, which is the operating-point effect of §4.1. Because the detector reads the image alone, and the same 50 MPCAttack images up-sampled to 299 px by bicubic interpolation and scored by the same frozen detector at the same thresholds are flagged on 0.00 against 0.98 at 224 px, every MPCAttack arm is scored at 224. The native sign-step MPCAttack and its TASS insertions at $\tau \in \{4, 8, 16\}$ are caught on 0.98, 0.98, 0.98 and 0.94, and inside the direct-pixel pipeline at step 2.5, where the realised rank stays near 196, TASS lowers detection of FTM, FOA-Attack, M-Attack and MF-ii only from 0.96 to 0.82, 0.76, 0.72 and 0.76 on the 50 scored ids (Table A1), consistent with the step-2.5 block of Table A2.

Basis controls. The compaction controls of §5.5 hold the budget and the pipeline fixed and change only the basis in which the perturbation is compacted, on 50 images at 10% false positives, with each arm's own success and energy. The singular basis reaches target top-1 0.76 at LPIPS 0.094 and RMS 4.93 and is flagged on 0.14 [0.06, 0.24]; a DCT basis reaches 0.77 at LPIPS 0.150 and RMS 6.21 and is flagged on 0.16 [0.06, 0.26], and a paired test between the two resolves no difference and establishes no equivalence. Two exploratory arms sit 14 to 16 points lower on success: pixel-domain compaction (0.62, LPIPS 0.066, RMS 3.54) is flagged on 0.12 [0.04, 0.22], and a random orthonormal basis (0.60, LPIPS 0.078, RMS 4.44) on 0.44 [0.30, 0.58]. The random arm is the informative one: at comparable compaction and energy it is flagged three times as often, so compaction as such is not what evades, and nothing here singles out the singular decomposition as the necessary basis.

Table A4: What the defender had to have seen. Every row trains on the same negatives and the same budget of attacked positives, 40 per fold split evenly over the families it may see (39 where three cannot divide 40), and takes its threshold from its own clean development pool (App. D). The scored arm is the compacted arm of Table 3a on its 50 scored images, with gains paired over them. A reduced-amplitude row multiplies the unconstrained attack's perturbation by 0.5 or 0.25 and re-exports it, giving realised RMS 6.41 and 3.27 against 12.80: copies of a known attack, not new attacks. The realised false-positive rates behind these rows run from 0.053 to 0.090 against a nominal 10%, so the operating points are close rather than identical.

Attacked images the detector trains on	TPR@10%	real. FPR	paired gain [95%]	on unconstr.
unconstrained attack only (zero-day)	0.26	0.090	–	0.98
+ the same attack at half amplitude	0.66	0.067	+0.40 [+0.26, +0.54]	0.96
+ the same attack at quarter amplitude	0.88	0.060	+0.62 [+0.48, +0.76]	0.96
+ both reduced amplitudes	0.88	0.053	+0.62 [+0.48, +0.76]	0.96
+ the compacted family itself	**0.94**	0.067	+0.68 [+0.56, +0.80]	0.96

The pair under a shared cap. A further control equalises the observed success and the L_2 cap, though not the realised energy: a free perturbation re-optimised inside the compacted arm's own cap reaches the same target top-1 of 0.76 at RMS 4.07 against 4.93 and effective rank 199 against 99, so the compacted arm is the one carrying more energy there. It is flagged on 0.14 $[0.06, 0.24]$ and the ball on 0.16 $[0.06, 0.26]$, a paired difference of -0.02 $[-0.06, 0.00]$ over the 50 shared images: no difference is resolved, and equivalence is not established. Because the two arms also differ in parameterisation and step size, this null is a null about two configurations rather than about the spectrum; the control that would isolate it, an L_2 ball or a low-pass constraint re-optimised in the direct-pixel skeleton at the rule's own step, we do not run.

Coverage protocol. Table A4 trains one detector per calibration setting under the frozen splits. Each row's negatives are the 300-image clean training pool and the clean counterparts of the 40 attacked ids outside the fold. Each row takes its threshold from its own 300-image clean development pool, because a different training set shifts the score distribution and a shared threshold would not deliver the nominal false-positive rate, and its realised rate is read on a further 300 clean images that no row trained on. The table's zero-day row is calibrated on the 500-pair launch of the unconstrained attack that Table 3a prints, whereas the zero-day detector of Table 3 itself calibrates on a 100-image launch of that configuration; the two agree at the 10% operating point, where both read 0.26, and differ by one image at 5%.

E GENERATIVE-VICTIM EVALUATION

Generative-victim protocol. The protocol is fixed in the released configuration before any victim is queried. *Pairs.* We exclude twelve of the 500 encoder pairs from the frame, because those images had already been shown outside the protocol. The remaining 488 pairs are sorted by the SHA-256 hash of the image id, and we take the first 200. *Arms.* The arms comprise the ten Baseline and TASS configurations of Table A1. The same table's ten HSS and USS configurations and the second attack seed of the Omni-Attack pair were added later, each before any query was made on it. The arms also include the clean image, the transfer version of AttackVLM (the native MF-ii pipeline), and the native M-Attack and MPCAttack images. As an exploratory follow-up added after Table 1 was seen, they further include the direct-pixel arm with no rule at step 2.5 on the Omni-Attack objective. Every arm submits its final evaluated image, and no victim contributes a gradient or a configuration choice. *Victims.* We query Qwen2.5-VL-7B-Instruct at a pinned revision with greedy decoding. We query `gpt-5.6-sol` (reasoning effort medium) and `gemini-3.8-flash` (thinking level medium) through their official APIs, and `qwen3.8-27b`, `deepseek-v4.1-flash` and `glm-5.3-flash` (temperature 0) through an institutional OpenAI-compatible endpoint. The victims were not all queried at once: four were queried over four consecutive days, with the rule groups spread over those days, and the remaining two in one later pass. Query dates therefore differ between the rules of one group by up to three days, a provider-side change within that window cannot be excluded, and every query date is in the released records. *Prompt.* We use "Name the main object in this image." in a fresh context with no target label, reference image, or method name. *Targeted ASR.*

A response counts only if it identifies the main object as the target class or as a registered equivalent from the class lemma list. A mention, a comparison, a negation, a list of alternatives, a secondary element of the scene, a hypernym alone, a refusal, or a technical failure does not count, and the denominator is always the full 200 pairs. Labels are first assigned by a rule-based matcher with the method names hidden. Two refinements to that matcher were made after inspecting responses, one concerning secondary-mention markers and one concerning the treatment of "corn/maize"-style lemma pairs, and both are recorded in the configuration and applied uniformly.

Label review. The blinded sheet, which replaces the method name by an opaque key, was re-labelled by an independent automated pass that was given the target class name, the registered synonyms, the response text and the matcher's own label, and was asked to decide each row under the scoring rule above; seeing the matcher's label is a limitation of the pass. The sheet holds the 21,200 responses of the four victims queried before it was frozen, and the pass returned a label on 18,000 of them. DeepSeek-V4.1 and GLM-5.3 were queried after the freeze and are not in the sheet. The two label sets agree on 89.4% of those rows, and the 1908 disagreements are asymmetric: on 1408 rows the matcher says no and the second pass yes, on 500 the reverse, so the second pass is the more permissive on balance, although 11 of the 106 arm–victim cells move down under it. Pooled over the covered rows the targeted rate is 25.6% under the matcher and 30.7% under the re-labelling, a mean shift of $+5.0$ points per cell with a range of -3.2 to $+23.2$. What the paper reads off these tables survives that shift: the clean-to-target control is 0.0% on every victim under both label sets, and 54 of the 60 re-checked baseline-to-rule comparisons keep their sign, the six exceptions being cells whose matcher difference was within 3 points and which the paper does not call resolved. This is a sensitivity check rather than a two-annotator arbitration: no human adjudication has been performed, an automated annotator that disagrees with a strict matcher is not by itself evidence that the matcher is wrong, and every printed ASR is therefore the matcher's, the conservative choice, with the re-labelling released alongside it. *Statistics.* We use a paired image bootstrap with 10^4 resamples over the 200 pairs, and we run it on one frozen attack seed. Every query records the image hash, the model identifier, the date, the prompt, the raw response, the label, and the retry history.

Generative victims: the table, intervals, seeds, and controls. The per-cell paired differences behind Table 1, each rule against its baseline configuration and TASS against USS, are in the released records. The clean controls are given here instead. Clean images name the target on 0.0% of the pairs on every victim, and they name the source class on 38.0% of them for Gemini-3.8-flash and 45.0% for GPT-5.6-Sol under the same strict rule, with the per-victim rates in the released records. The transfer version of AttackVLM names the target on 0.0% on every victim and the source class on 31 to 38%. Every printed rate is the rule-based matcher's, and the label review above reports the independent re-labelling and how far the rates move under it. The table below gives the main pair at attack seed 1. The seed-0 differences are in the released records.

Victim	Baseline ASR (%)	+ TASS ASR (%)	Paired difference (points)
GPT-5.6-Sol	27.0	33.0	$+6.0$ $[+0.5, +11.5]$
Gemini-3.8-flash	37.5	34.5	-3.0 $[-10.5, +4.5]$
Qwen3.8-27B	43.0	43.5	$+0.5$ $[-6.0, +7.0]$
DeepSeek-V4.1	39.0	44.5	$+5.5$ $[-1.0, +12.0]$
GLM-5.3	32.5	43.0	$+10.5$ $[+3.0, +18.0]$

The stealth-regime arm on the open-weight victims. Table 1 is read in the high-transfer regime, so the stealth-regime TASS arm of the Omni-Attack objective (τ=1, step 0.5) was also sent to the four open-weight victims on the same 200 pairs, prompt and scoring rule. It reaches 32.0, 43.0, 41.0 and 35.0 on Qwen2.5-VL, Qwen3.8-27B, DeepSeek-V4.1 and GLM-5.3 against the baseline configuration's 36.5, 43.5, 40.0 and 35.0, and the paired difference covers 0 on all four. Against the high-transfer TASS arm it is a resolved loss on two of the four, -10.5 $[-16.5, -5.0]$ on Qwen2.5-VL and -9.0 $[-17.0, -1.0]$ on GLM-5.3, and unresolved on the other two, so the regimes differ in transfer and we do not claim they are equivalent. The two closed-source victims were not queried in this regime, and detection is read on the encoders throughout.

Generative control on the main objective. Table A5 asks whether the generative-victim differences of Table 1 can be explained by the direct pixel update or by uniform shrinkage. Because the

Table A5: Exploratory generative-victim control on the Omni-Attack objective at step 2.5 on the 200 frozen pairs, added after the main results were seen. The rows share pairs, prompt, victims, settings, and attack seed with the main evaluation, and the USS threshold is the pre-existing nearest-RMS point. Every row but the baseline, which runs at step 0.05, is a direct-pixel arm.

	Targeted ASR (%), and paired difference in points						
Configuration	GPT-5.6	Gemini-3.8	Qwen2.5-VL	Qwen3.8	DeepSeek	GLM-5.3	LPIPS
Baseline ($\epsilon \tanh$)	26.5	32.0	36.5	43.5	40.0	35.0	0.346
None (τ=0)	31.0	35.0	41.5	42.0	44.5	41.5	0.259
USS (τ=1)	27.5	37.0	42.0	45.0	41.5	39.0	0.238
TASS (τ=1)	31.0	35.0	42.5	46.5	47.5	44.0	0.248
TASS minus None	$[-5.0, 5.0]$	$[-6.5, 6.5]$	$[-5.0, 7.0]$	$[-1.5, 10.5]$	$[-2.5, 8.5]$	$[-4.5, 9.5]$	
TASS minus USS	$[-1.5, 8.5]$	$[-9.0, 5.0]$	$[-5.5, 6.5]$	$[-4.5, 7.5]$	$[-0.5, 12.5]$	$[-1.0, 11.0]$	

Table A6: Native configurations of the external attacks and the full Omni-Attack + TASS configuration on the same 200 pairs, budget, and generative victims. Each row uses its own surrogates and step count, listed here.

			ASR (%)						
Method	Surrogates	Steps	GPT-5.6	Gemini-3.8	Qwen2.5-VL	Qwen3.8	DeepSeek	GLM-5.3	LPIPS
AttackVLM (MF-ii)	B/32	300 sign	0.0	0.0	0.0	0.0	0.0	0.0	0.314
M-Attack	B/16, B/32, g/14	300 sign	3.0	12.0	12.5	14.0	13.0	15.0	0.304
MPCAttack	3 CLIP, InternVL, DINO	300 MI-FGSM	6.5	27.5	21.5	19.5	23.0	23.5	0.372
Omni-Attack	5 CLIP	1500 Adam	26.5	32.0	36.5	43.5	40.0	35.0	0.346
Omni-Attack + TASS	5 CLIP	1500 Adam	31.0	35.0	42.5	46.5	47.5	44.0	0.248

Source: vestment → Target: corn Prompt: "Name the main object in this image."

Configuration		LPIPS	GPT-5.6-Sol	Hit	Gemini-3.8-flash	Hit
Clean			A group of clergy in purple robes.	×	The main subject in the image is a group of altar servers (or altar boys) standing alongside a […]	×
Baseline		0.325	A group of people.	×	Based on the composite/layered image, the main subjects visible are a group of people sitting […]	×
+HSS		0.207	Purple **corn** cobs.	✓	Based on the image, the main objects depicted are ears of **corn** (**corn** cobs) blended or overlaid […]	✓
+USS		0.228	Yarn skeins.	×	The main objects depicted in this hybrid/surreal image are ears of **corn** (**corn** cobs) blended […]	✓
+TASS		0.223	**Corn** on the cob.	✓	The main object depicted in this image is **corn** (ears of purple and white corn/maize), surrealistically […]	✓

Figure A2: Recorded responses to one source image under the clean reference and the four Omni-Attack configurations; both models receive the same file and neutral prompt per row, a tick marks a target hit, and LPIPS is against the clean image; one pair establishes no ordering.

run was carried out only once those results were available, we read it as exploratory. The paired differences in the table are the evidence, and no threshold was reselected on these victims.

Native comparison on generative victims. Table A6 sends the native images to the same victims. The native attacks reach 0.0 to 27.5 ASR from their own surrogates. The strongest of them is MPCAttack, and Omni-Attack is above it by 20.0 points on GPT-5.6-Sol, 4.5 on Gemini-3.8-flash, 15.0 on Qwen2.5-VL, 24.0 on Qwen3.8-27B, 17.0 on DeepSeek-V4.1 and 11.5 on GLM-5.3. These rows show what each released configuration does on this task under the protocol we evaluate. They are not evidence about those methods on their own benchmarks.

Figure A3 shows the next pair in the frozen hash order as a response matrix. On that pair both models name the target under the baseline and under TASS, while LPIPS falls from 0.187 to 0.117,

Source: triumphal arch → Target: electric fan Prompt: "Name the main object in this image."

Configuration		LPIPS	GPT-5.6-Sol	Hit	Gemini-3.8-flash	Hit
Clean			Arc de Triomphe	×	The main object in the image is the Arc de Triomphe (Arc de Triomphe de l'Étoile) in […]	×
Baseline		0.187	**Electric fan**	✓	The main object in the image is an **electric fan** (pedestal/standing fan).	✓
+HSS		0.108	**Electric fan**	✓	The main object in the image is a vintage **electric fan.**	✓
+USS		0.119	**Electric fan.**	✓	The main object in the image is a vintage **electric fan.**	✓
+TASS		0.117	**Electric fan.**	✓	The main object in this image is a vintage **electric fan** (specifically, an antique General Electric […]	✓

Figure A3: Recorded responses to a second source image, in the layout of Figure A2: both models receive the same file and neutral prompt per row, a tick marks a target hit under the frozen rule, and LPIPS is against the clean image; one pair establishes no ordering.

which is the reduced-distortion case. The albatross pair is a miss for both attacks on both models, and we keep it in the released records rather than showing it as a figure.

F External baselines and scope

Ported losses. FTM (Liang et al., 2025) perturbs the surrogate's intermediate features with a learnable tensor that is optimised jointly with the image, and it then mixes those features stochastically with the clean ones (β=0.01, update probability 0.1, $\alpha \sim U(0, 0.75)$). FOA-Attack (Jia et al., 2025) aligns to a target image in two ways at once. It aligns globally on the `[CLS]` token, and it aligns locally by optimal transport between the K-means centres of the patch tokens (n=5, λ=0.1), while each surrogate is weighted by its rate of loss decrease. M-Attack (Li et al., 2025) matches the cosine to a randomly cropped target image, where the crop scale runs from 0.5 to 0.9, and MF-ii (Zhao et al., 2023) matches the cosine to the uncropped target image. Each of these losses runs inside our pipeline with our surrogates, augmentations, and optimiser, so the rows of Table A1 are not reproductions of the authors' headline numbers.

Native pipelines and the sign-step boundary. We also ran M-Attack, MF-ii and MPCAttack in their own pipelines, with sign steps, their own surrogates, 224 px and 300 steps, on our 500 pairs and victims: M-Attack reaches top-1 34.7, MF-ii 0.0, and MPCAttack, which optimises an InfoNCE loss over three CLIP encoders, InternVL and DINOv2 with MI-FGSM, 56.2. Inserting TASS into those unmodified sign-step updates does not raise success. For MPCAttack, τ=8 gives 55.6, a change of -0.6 $[-2.4, +1.2]$, at LPIPS 0.348 against 0.363, with τ=4 and τ=16 at 53.7 and 53.3; for M-Attack, $\tau \in \{2, 4, 8\}$ lowers top-1 by 7.6 to 20.6 points. We state this only for the sign-step pipelines we tested. Under a $\pm$1-valued step that replenishes the spectrum at every iteration the shrinkage does not compact the iterate: the arms stay near effective rank 160 whatever τ is set to, and the zero-day detector flags them (App. D). Replacing MPCAttack's sign step by direct-pixel Adam, with step 0.5, 3000 steps and 224 px, inside its own code and surrogate pool gives the pipeline of Table A3. Over the 500 pairs it reaches top-1 44.3 with no rule, below the native 56.2, and TASS at τ=0.25 is not distinguished from it, at $+1.4$ $[-1.7, +4.4]$ (Table A3 prints the 50 scored ids), at LPIPS 0.175 against 0.372 and effective rank 74 against 170, whereas τ=0.5 and τ=1 cost 4.0 $[0.8, 7.3]$ and 10.2 $[7.0, 13.4]$ points. Because τ is an absolute threshold and this pipeline's singular values are about five times smaller than those of the step-2.5 arms, the same nominal τ is a far stronger operating point here, which is the mechanism of §4.1.

Attacks not compared. Our re-implementation of V-Attack (Nie et al., 2026) moves the image toward the target class by only $+0.011$ cosine on 25 images and never takes the top-1 slot; because its authors' pipeline differs from ours in optimiser and in surrogate family, we record this as a limitation of our re-implementation rather than as a result about the method. AnyAttack (Zhang et al., 2025), a generator pre-trained on LAION-400M, reaches 0.1 targeted top-1 on our cohort while still moving images in its own CLIP ViT-B/32 space, and it supplies no gradient for a TASS counterpart. Universal perturbations (Huang et al., 2025; Xu et al., 2025) address a different threat model.

G Generalisation across budgets, datasets and victims

Exploratory non-inferiority across budgets. On panel (a) of Table A7 a one-sided check requires the upper 95% bootstrap bound of the baseline-minus-TASS difference to lie below a 2-point margin. It holds for top-1 at both budgets, where the intervals are $[-2.1, +1.5]$ at ϵ=16 and $[-6.4, -0.6]$ at ϵ=8, and for the weaker tvg criterion at ϵ=16, where the levels are 98.4 against 97.9 and the interval $[-0.2, +1.3]$; it does not hold for tvg at ϵ=8, where the levels are 89.0 against 88.0 and the interval $[-0.5, +2.5]$. The margin was fixed after inspecting the data, so the check supports the equal-success reading without being a confirmatory test.

Second dataset and supervised victims. Table A7 holds the numbers behind §5.6. The ImageNetV2 cohort is matched-frequency and contains 500 images, and it shares the label space, the surrogate and the victim split of the main collection. On that cohort the step-0.5 TASS ties Omni-Attack+L_2 ($+0.8$ $[-0.8, +2.4]$) and Omni-Attack (-0.1 $[-1.6, +1.4]$) on top-1 at effective rank 125, whereas the step-2.5 TASS exceeds both ($+3.0$ $[+1.2, +4.8]$ and $+2.1$ $[+0.3, +4.0]$) at rank 195. The four-variant group at step 2.5 reproduces the picture of the main collection, with TASS minus no rule at -0.1 $[-1.5, +1.3]$, TASS minus USS at -0.4 $[-1.7, +0.9]$ and TASS minus HSS at -0.1 $[-1.7, +1.4]$, and with HSS again at the lowest LPIPS among those four variants (0.225).

Table A7: Generalisation: two budgets on NIPS-2017, a second dataset, and supervised victims. Each panel has its own success metric, cohort and seed count, so comparisons are made within a panel and never across them. The ϵ=16 rows of panel (a) are a separate launch from Tables A1 and A2 and differ from them by a few tenths.

(a) Budgets on NIPS-2017, 500 *pairs, target top-1;* ϵ=8 *over* 3 *seeds,* ϵ=16 *one seed*

Budget	Method	Top-1 (%) ↑	LPIPS ↓	RMS ↓
ϵ=8	Omni-Attack	60.3	0.161	6.69
	TASS (τ=1, step 0.5)	63.8	0.094	4.44
ϵ=16	Omni-Attack	85.1	0.339	12.73
	TASS (τ=1, step 0.5)	85.4	0.183	7.36

(b) ImageNetV2, 500 *images, target top-1, one seed*

Method	Top-1 (%) ↑	LPIPS ↓	RMS ↓	Eff. rank
Omni-Attack	88.1	0.341	12.70	221
Omni-Attack+L_2 (ball 7.33, energy control)	87.2	0.163	–	208
Direct pixel, no rule (step 2.5)	90.3	0.252	9.45	–
HSS (r=128, step 2.5)	90.3	0.225	8.63	–
USS (τ=1, step 2.5)	90.6	0.229	8.54	–
TASS (τ=1, step 2.5)	90.2	0.237	8.82	195
TASS (τ=1, step 0.5)	88.0	0.176	–	125

(c) Supervised CNN victims, 500 *images, target-over-ground-truth win rate, one seed*

Method	ResNet-50 (%)	ConvNeXt-B (%)
Omni-Attack	69.2	77.4
TASS (τ=1, step 0.5)	58.8	69.4

The least distorted arm on this cohort is Omni-Attack+L_2, which reaches 0.163 against 0.176 for low-rank TASS, a paired difference of -0.013 $[-0.015, -0.011]$, so no distortion ordering across families is claimed. Only one ball radius was run, so this replicates two operating points rather than a frontier. Panel (c) reports the win rate of an independently optimised n=500 run, where TASS trails Omni-Attack on ResNet-50 and on ConvNeXt-Base. The CNN metric is not comparable with target top-1.

H Additional visual examples

Figures A4–A6 are selected after the fact and on the outcome, which the blind draw of Figure 2 is not: they are pairs the hardest closed-source victim gets wrong. Among the frozen 200 of App. E the rule keeps a pair when GPT-5.6-Sol names the target class under at least 20 of the 21 objective-by-rule arms queried on it (the no-rule arms of the four objectives other than Omni-Attack were not queried), does not name it on the clean image, has at most 2% of clean pixels saturated at 0 or 255, and has a saved image in every arm; pairs are then ordered by hit count and by id. That rule returns exactly three pairs, so no further choice was made, and the released script recomputes it rather than storing the answer. These figures therefore show what a successful attack looks like, not what a typical one looks like: the per-victim rates are in Table 1, and 102 of the 200 pairs move GPT-5.6-Sol at all. Within a figure, each crop window is the lowest-variance window of that clean image on a 16-px grid, fixed before any attacked image was read. No figure establishes an ordering, the aggregate metrics use the full evaluation set, and the images are the saved evaluated PNGs: nothing is recoloured, denoised, or regenerated.

Figure 2, by contrast, is drawn blind. The 50-image pool of the released selection record, minus the six images whose clean pixels are more than 2% saturated at 0 or 255, is ranked by clean-image texture (mean gradient magnitude) and split into three strata, one id is drawn per stratum with seed 0, and the figure shows the middle-stratum draw; its window is the lowest-variance 64×64 window of the clean image on the same 16-px grid. The released script replays the draw rather than storing its answer.

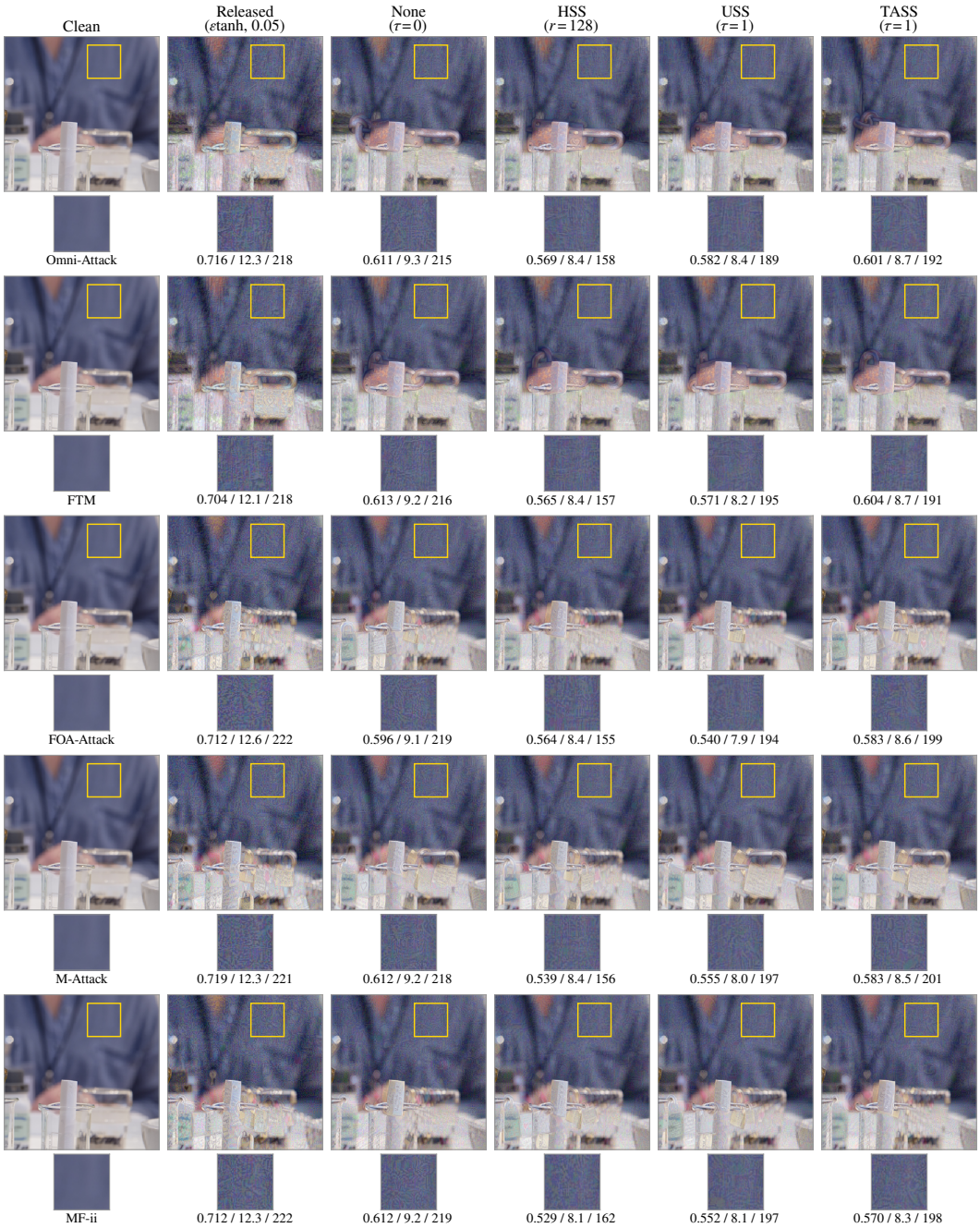

Figure A4: A pair GPT-5.6-Sol misidentifies, under all five attack objectives at the settings of Table A1. The clean image is labelled *beaker* and the attacker's target class is *padlock*; GPT names the target under all 21 of the arms queried on it and names it on no clean image. Each block is an objective: the image row, then the same 64×64 window magnified. Labels give LPIPS, per-pixel RMS and effective rank.

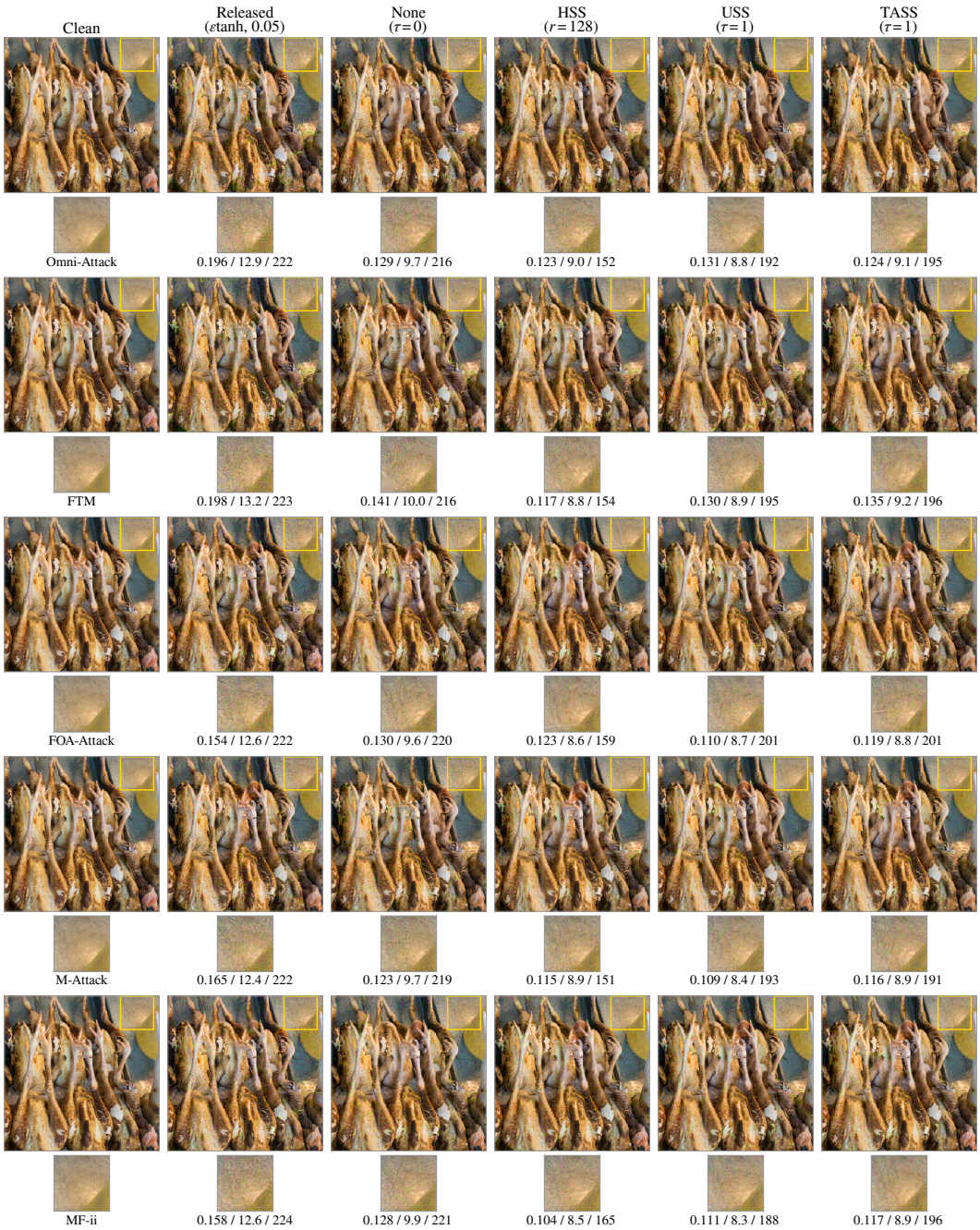

Figure A5: A pair GPT-5.6-Sol misidentifies, under all five attack objectives at the settings of Table A1. The clean image is labelled *butcher shop* and the attacker's target class is *gorilla*; GPT names the target under 20 of the 21 arms queried on it and names it on no clean image. Each block is an objective: the image row, then the same 64×64 window magnified. Labels give LPIPS, per-pixel RMS and effective rank.

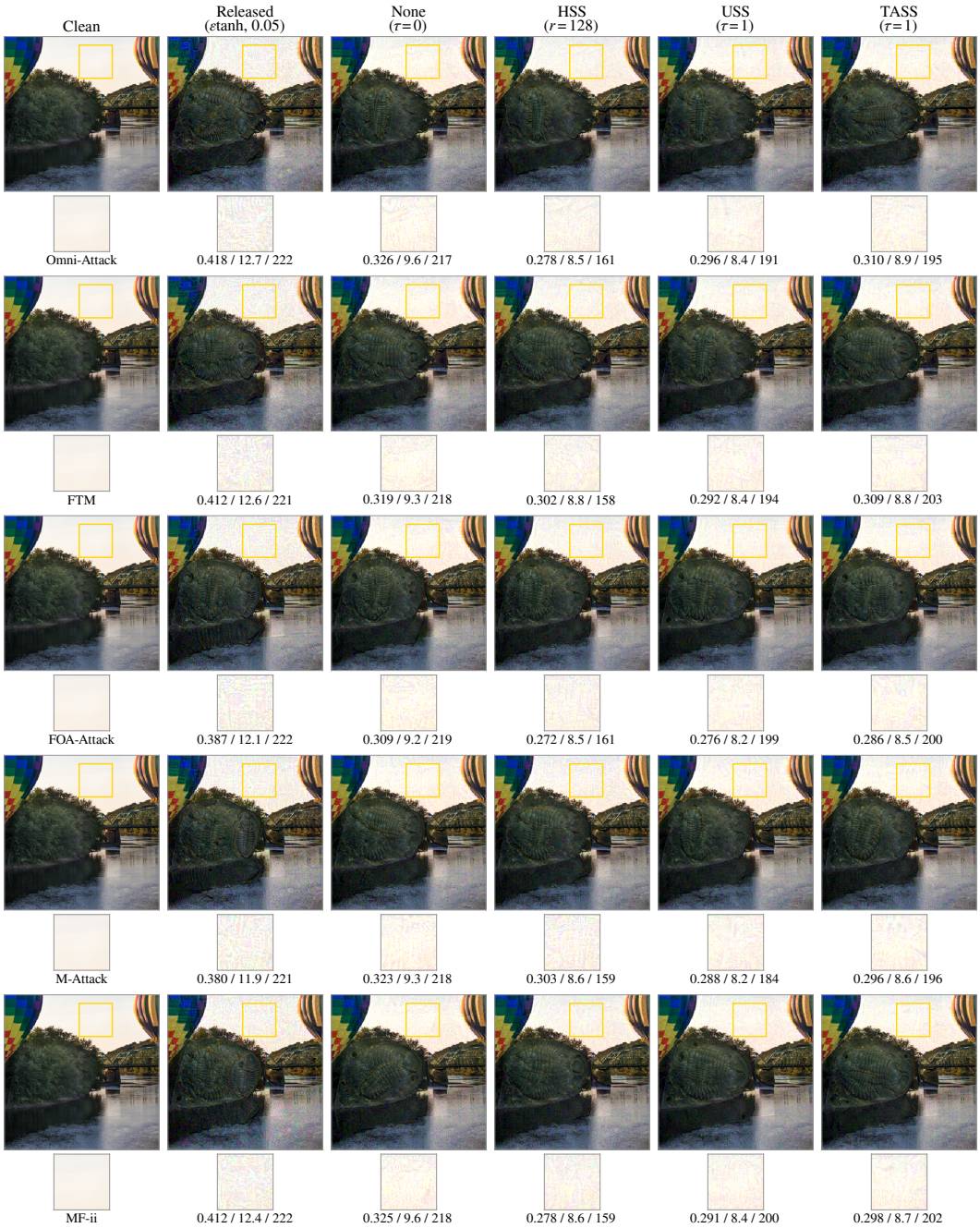

Figure A6: A pair GPT-5.6-Sol misidentifies, under all five attack objectives at the settings of Table A1. The clean image is labelled *balloon* and the attacker's target class is *trilobite*; GPT names the target under 20 of the 21 arms queried on it and names it on no clean image. Each block is an objective: the image row, then the same 64×64 window magnified. Labels give LPIPS, per-pixel RMS and effective rank.