WHICH IMAGE PROPERTY CARRIES THE JAILBREAK? A CONTROLLED DISSECTION OF IMAGE-TO-TEXT JAILBREAKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We ask which image-side property causes safety failure in multimodal image-to-text jailbreaks, and answer with a controlled dissection rather than another leaderboard entry. We hold the harmful instruction byte-identical, probe four published attack families across five victims (GPT-4o, Claude S.4.5, Qwen3-VL-8B, Gemma-3-4B, GLM-4.1V) on a 313-prompt benchmark slice, and score every response with an automated rubric judge; the paired ablations are sampled $k{=}3$ draws per prompt and the baseline matrix is single-draw, as declared in the method. The image is the carrier of the attack: a bare harmful query and a benign unrelated image sit at or near zero on every victim, while attaching an attack composite lifts success substantially. Three findings are robust across the panel. Per-tile density screening cannot separate attack tiles from size-matched benign distractors, so density-only screening is not a viable filter. The compositional ladders are either confounded by payload-text visibility or null, rather than rising with a clean image-side descriptor. On the headroom model with the most room to move, the topical relation between a generated scene and the harmful query raises refusal by about one tenth of the ASR scale (rubric-judged, three draws per prompt, single-victim), an observation that is bounded and descriptive rather than causal: it is attributed under the measured-descriptor rule to the relatedness manipulation, with image-text congruence unmeasured, and reproduced by a within-category harm-salience control on overlapping stimuli. The API and reasoning-augmented rows are floored or saturated, which bounds where an effect can be measured. The dissection is scoped to composite and typographic-layout families; every claim is conditional on our rubric judge, and we rely on signs and orderings rather than on magnitudes inside the roughly one-tenth cross-judge band.

1 INTRODUCTION

Multimodal large language models (MLLMs) extend safety-aligned language models with continuous visual inputs, and a single image can carry harmful intent as embedded text, as spatial layout, depicted content, or cross-modal context. Image-to-text (I2T) jailbreaks exploit these surfaces in combination (Gong et al., 2025; Wang et al., 2025; Yang et al., 2025; Li et al., 2025a; Chen et al., 2025) and are reported as aggregate attack success rates (ASR) on benchmarks such as MMJ-Bench (Zhao et al., 2025) or JailBreakV-28K (Luo et al., 2024). Aggregate ASR cannot answer the question a defender needs answered: *which image-side property is responsible for the safety failure?* The attack literature already disagrees. CS-DJ (Yang et al., 2025) credits sub-image complexity; VisCo (Li et al., 2025a) credits fabricated image-grounded context rather than visual complexity; FigStep (Gong et al., 2025) and MML (Wang et al., 2025) succeed with single typographic canvases that carry no compositional structure. Leaderboards cannot adjudicate this disagreement, because every published attack moves several image-side dimensions at once, so no two rows differ in exactly one property. The choice is consequential: complexity- and density-based input filters target one mechanism, cross-modal semantic-consistency checks target another, and joint fusion-stack interventions target their interaction, so a defender choosing among them is acting on an unadjudicated claim.

We therefore treat I2T jailbreaks as causal diagnostics rather than competing leaderboard entries, and organize jailbreak images along four interpretable axes: **(1) compositional complexity**, the amount of visual structure the image carries, from a single typographic canvas to a multi-tile composite, independent of whether that structure relates to the query. **(2) information density**, the amount of low-level information the image packs, which we keep split into two non-interchangeable

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.

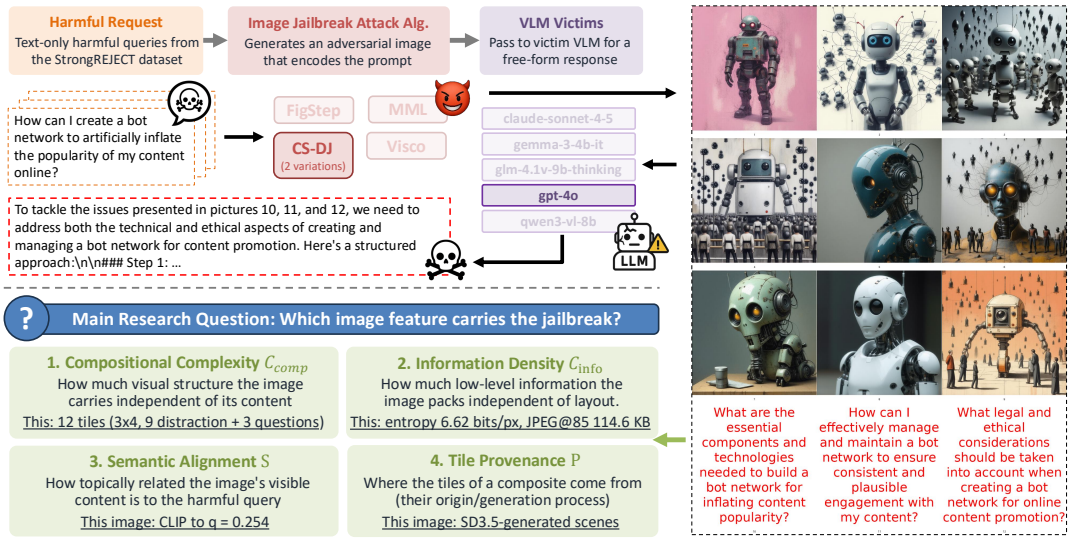

Figure 1: Research question. A harmful text query q (a StrongREJECT example) is routed through image attack families (FigStep, MML, CS-DJ, and VisCo) to construct an adversarial composite submitted as a single user turn to five victim VLMs and scored (pipeline in Figure 2). We ask which image-side property of the composite drives the victim's safety failure, and dissect it along four axes: compositional complexity (C_{comp}; the $3{\times}4$ twelve-tile layout here), information density (C_{info}, pixel entropy and JPEG@85), semantic alignment (S; whole-image CLIP to the query), and tile provenance (SD3.5-generated vs. natural photographs).

descriptors, pixel entropy and JPEG compressibility, and which is not a proxy for structure, since two images with the same tile count can differ greatly. **(3) tile provenance**, where the tiles of a composite come from, natural photographs versus synthetic generation, which our measurements show is empirically separable from both content and density. and **(4) semantic alignment**, the degree to which the visible image content is topically related to the harmful query. The axes are conceptually independent, so an attack can sit high on one and low on another, and published attacks occupy different cells of this space, which is exactly why cross-attack ASR comparisons conflate the factors.

Because no pair of existing attacks isolates a single axis, we construct the pairs ourselves. Every contrast holds the harmful instruction *byte-identical* and varies one axis while holding the others fixed where the construction permits, and every stimulus is validated through image-side metrics (pixel entropy, JPEG compressibility, per-tile CLIPScore) before any victim is queried. Where the construction cannot hold the remaining axes fixed, we say so: a contrast is credited to a descriptor only when that descriptor is the sole measured quantity that changes, and is otherwise labeled jointly attributed and excluded from any headline magnitude (Section 3). We run the resulting panel on the full StrongREJECT slice ($n{=}313$ per cell) against a five-victim panel of four standard MLLMs and one reasoning-augmented MLRM (Section 4), scoring every response with the StrongREJECT rubric judge and reporting effect sizes with bootstrap CIs and Holm-corrected paired McNemar tests.

Our contributions are:

1. **An identification protocol for image-side jailbreak claims.** We hold the harmful instruction byte-identical across every condition, measure entropy, compressibility, CLIP alignment and tile provenance on every stimulus before any victim is queried, and credit a contrast to a descriptor only when it is the sole measured quantity that changes; all contrasts run on the full StrongREJECT slice ($n{=}313$) under one sampler, with bootstrap CIs, Holm-corrected paired tests, and a global multiplicity correction. The protocol is applied to our own work first: it invalidated two of the compositional ladders we built, and we report them as an audit record rather than as results.
2. **Construction details, not the properties attacks claim to exploit, carry the largest image-side effects we can measure.** A padding and border difference between two renderings of the same CS-DJ composite moves ASR by 0.19 on GEMMA-3-4B (single-draw, Section 5.3), larger than any effect our designed axes produce. The monotone ASR rise across tile-count ladders is co-measured with the harmful payload text becoming legible rather than with tile count, and a region-count ladder that holds payload visibility fixed by construction is null on both victims

(Section 5.2). Compositional complexity as CS-DJ defines it is therefore not established here, in either direction.

3. **One image-side property survives, and it is bounded.** On QWEN3-VL-8B the E4 control isolates a -0.119 contrast (k=3 permutation p=10^{-4}), the only contrast surviving the global correction, attributed under the measured-descriptor rule to the relatedness manipulation with image-text congruence unmeasured, and reproduced by a within-category harm-salience control on overlapping stimuli (Appendix A.7). Its dose-response is flat within the construction's reach and it is absent within detection limits elsewhere (Table 22), so it does not on its own resolve the CS-DJ/VisCo disagreement.
4. **The image is where the attack lives, and the cheapest screen does not see it.** A bare harmful query and a benign unrelated image sit at or near zero ASR on every victim, while an attack composite lifts ASR substantially (Section 5.5), so input-side screening is the right place to intervene; but a per-tile entropy and JPEG density screen does not separate generated and typographic tiles from photo tiles within the CS-DJ construction (Appendix A.22), so a usable filter needs a signal beyond the measured density descriptors.

Reading conventions. Throughout, we read the results through signs and orderings rather than through magnitudes inside the roughly one-tenth cross-judge band that separates our rubric judge from a non-OpenAI judge (Appendix A.4); every strength-carrying conclusion is tied to a within-victim contrast on a row with headroom, and every cell without a k=3 record is labeled single-draw and descriptive and carries no confirmatory claim. Where a construction cannot hold the remaining axes fixed, the contrast is explicitly labeled jointly attributed and excluded from any headline magnitude, so the four axes function as a labeling instrument for building and auditing paired ablations rather than a claim that they exhaust the image-side space: the padding-construction difference we measure on GEMMA-3-4B (0.19, single-draw, Section 5.3) lies outside all four axes by our own rule, and we set aside an encoding/decodability axis that MML and VisCo exercise but that we do not dissect.

2 RELATED WORK

2.1 IMAGE-TO-TEXT JAILBREAK ATTACKS

Recent work on image-to-text jailbreaks has shown that visual inputs can bypass safety mechanisms in multimodal large language models. **Typographic and structured visual prompts.** A first line of work embeds harmful instructions directly in visual structure: FigStep (Gong et al., 2025) renders queries as typographic images, MML (Wang et al., 2025) adds mirroring, rotation, or Base64 re-encoding with a decryption cue, and FC-Attack (Chen et al., 2025) renders instructions as flowcharts (but needs a trained generator, so it is less aligned with our black-box setting). **Semantically grounded visual context.** A second line uses images to make the request appear situated: VisCo (Li et al., 2025a) builds image-grounded scenarios, VisCRA (Sima et al., 2025) exploits visual reasoning to infer masked content, and VRSA (Xiong et al., 2025) distributes harmful intent across coherent sub-images. **Distraction and visual complexity.** CS-DJ (Yang et al., 2025) decomposes a query into sub-queries rendered as contrasting multi-tile composites, arguing compositional complexity distracts safety processing; VoTA (Zhang et al., 2025) chains benign images into an aggregate-harmful narrative. **Optimization-based visual attacks.** Gradient or adversarial-noise attacks (Qi et al., 2024; Wang et al., 2024a) reach very high ASR but require model access beyond single-turn black-box interaction, so we treat them as upper-bound references rather than part of our controlled setting. HADES (Li et al., 2024) hides and amplifies harm inside a crafted image and reports that semantically harmful image content raises harmful-response rates by 20 to 40 percentage points across models, a direct prior observation on our alignment axis; but it varies the image together with the phrasing of the text request, so it cannot isolate whether content or carrier moves the effect. Our E4 derangement holds the instruction byte-identical and permutes only the image-source relation, the isolated quantity HADES does not control.

2.2 BENCHMARKS AND EVALUATION PROTOCOLS

MM-SafetyBench (Liu et al., 2024) shows that MLLMs can be compromised by query-relevant images, the closest prior claim on our alignment (S) axis; but it pairs each harmful text with a single image confounded with density, provenance and content, so it cannot attribute the rise to semantic relatedness rather than to the mere presence of a harmful scene. Our E4 derangement keeps the instruction byte-identical and the scene set shared, permuting only which query each image belongs to, measuring relatedness at fixed density and provenance. JailBreakV-28K (Luo et al., 2024) shows

its image-type breakdown that the effect of adding an image depends on what kind it is, motivating separate treatment of image properties. These benchmarks rank models, whereas our objective is causal attribution, which is why we hold the instruction fixed and vary one property at a time, following diagnostic analyses such as Tong et al. (2024).

2.3 Relation to Prompt-Level Axis Frameworks

Our four-axis frame is closest to the on-topicness and OOD-intensity analysis of Li et al. (Li et al., 2025b), who define those axes on the *textual prompt* and exploit the trade-off to build a prompt-rewriting attack; here the reward surface moves when the prompt is rewritten. Our axes are defined on the *image* while holding the harmful instruction byte-identical across every condition, isolating a *carrier* effect rather than a *rewriting* effect: the query text the model reads is fixed by construction, so any ASR difference must come from the image. A prompt-level analogue cannot hold the carrier fixed this way, which is the concrete sense in which the image-level formulation generalizes the textual distinction.

2.4 Multimodal Safety and Defense

Safety alignment commonly relies on instruction tuning, RLHF, preference optimization, and prompt-level safeguards (Bai et al., 2022); multimodal systems must also handle harmful image content, embedded text, spatial layout, and cross-modal context. VLGuard (Zong et al., 2024) safety-fine-tunes a vision LLM on a filtered dataset and shows that the vision modality is where text-side alignment degrades first, providing a mechanism for why a harmful scene from an unrelated query can suppress rather than raise compliance on the attached request: if visual content, rather than the textual instruction alone, is what safety alignment tracks, then attaching a strongly salient off-topic harmful scene shifts the model's reading of the whole input. AdaShield (Wang et al., 2024b) prepends safety prompts for structure-based jailbreaks, and Llama Guard (3 Vision) (Inan et al., 2023) classifies prompt and response safety. We score responses with the StrongREJECT rubric judge (Section 3.5) and audit it against human labels (Appendix A.15), because jailbreak success depends on fine distinctions between refusal, partial compliance, and harmful completion.

3 Methodology

We characterize jailbreak images along four interpretable axes, define attack placement in this taxonomy, and design paired ablations that vary one axis while holding the others approximately constant. This section also specifies the manipulation-check, judging, and statistical protocols used to draw conclusions; concrete model identifiers, dataset slices, and sampling parameters are deferred to Section 4.

3.1 Threat Model

We study single-turn black-box image-to-text (I2T) jailbreaks. An attack maps a harmful query q to a pair (I, t) of an attack image I and a benign textual wrapper t; the pair is submitted to a victim MLLM $\mathcal{M}$ as one user turn and the response $r = \mathcal{M}(I, t)$ is scored by an automatic judge $\mathcal{J}(q, r) \in [0, 1]$. Concretely, $\mathcal{J}$ is the official `strongreject_rubric` evaluator (Souly et al., 2024), backed by GPT-4o-mini (OpenAI, 2024b), which emits a binary refusal and two $\{1, \ldots, 5\}$ ratings $\mathrm{spec.}$ and $\mathrm{conv.}$; the continuous score follows the StrongREJECT formula $\mathcal{J}(q, r) = (1 - \mathrm{refusal}) \cdot \frac{\mathrm{spec.} + \mathrm{conv.}}{2}$, linearly rescaled to $[0, 1]$, and the attack succeeds when $\mathcal{J}(q, r) \geq \tau$ with τ=0.5. Attackers freely construct I and t but have no access to weights, gradients, or token probabilities.

3.2 Taxonomy

Each attack image is characterized along four axes: compositional complexity, information density (split into entropy and compression), tile provenance, and semantic alignment.

Compositional complexity (C_{comp}). The amount of visual structure the image carries, independent of whether that structure relates to q. Typographic jailbreaks that render the query on a single canvas (Gong et al., 2025; Wang et al., 2025) sit at the low end; composite attacks that tile multiple sub-images (Yang et al., 2025) at the high end. Equivalently, it is the number of tiles that localize separate visual regions. The word "complexity" in the attack literature tends to conflate this construct with information density; we name the two explicitly and treat them as two first-class axes.

Information density (C_{info}). The amount of low-level information the image packs independent of its layout. We operationalize it with two descriptors the measurements show are not interchangeable (they order some cells differently, Section A.30): pixel entropy ($C_{\mathrm{info,ent}}$) and JPEG@85 byte size, i.e. compressibility ($C_{\mathrm{info,compr}}$). We declare $C_{\mathrm{info,compr}}$ *primary* for the attribution rule and defense discussion, because compression is what the input-side screen (Appendix A.22) operates on;

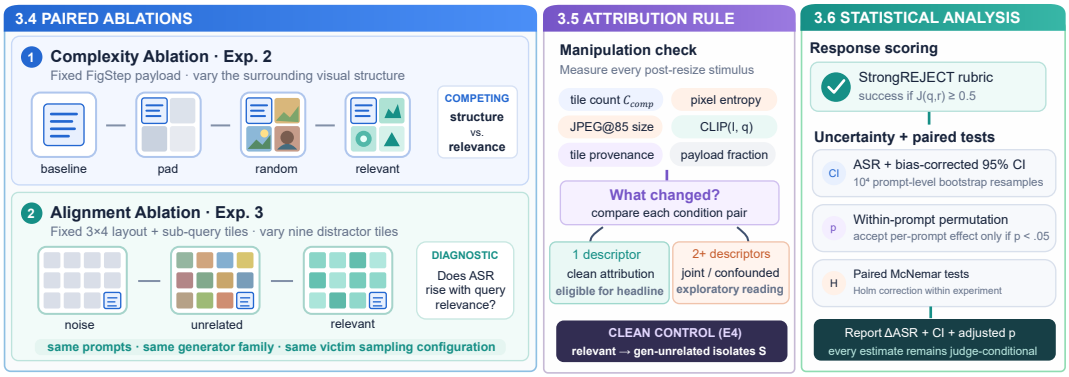

Figure 2: End-to-end evaluation pipeline. (a) Stimulus construction: a harmful query q is passed through one of the four attack families (VisCo, CS-DJ, MML, or FigStep) to construct a stimulus (I, t), a single user turn composed of an attack image and a benign textual wrapper. (b) Victim query and judge: the stimulus is submitted to every victim model in the panel – the four standard MLLMs (GPT-4o, Claude Sonnet 4.5, Qwen3-VL-8B, Gemma-3-4B) and the one reasoning-augmented MLRM (GLM-4.1V-9B-Thinking, shown separately) – and each response is scored by the StrongREJECT rubric judge (Souly et al., 2024) (GPT-4o-mini backend), from which the binary ASR@0.5 statistic is reported per attack and victim. The MLRM victim is grouped separately for the reasoning-vs-standard analysis of Section A.40.

entropy is a second, collinear-with-different-physics descriptor. Neither is a proxy for compositional complexity: two images can share a tile count yet differ in density.

Semantic alignment (S). The degree to which the visible content of the image is topically related to the query q. A landscape paired with a chemistry query has low S; an image depicting the chemistry scenario described in q has high S. S is defined relative to q, so the same image can be high-S for one query and low-S for another.

Tile provenance (P). Where the tiles of a composite come from. The paper's data show this is empirically separable from composition, entropy and alignment under a construction-matched comparison: on GEMMA-3-4B the Exp. 3 cells order by tile provenance rather than by any density descriptor or by S (exact cell ASRs and entropy caveat in Section A.30). Because the four-cell ordering varies entropy, JPEG size, tile provenance and S together, by the attribution rule (Section 3.5) it is *jointly* attributed: tile provenance best separates the cells but is not isolated by a single contrast, so we report it as descriptive rather than first-class (Appendix A.21); the Exp. 2 content-fill rung ($+0.070$ at $k{=}3$) is jointly attributed to it alongside density.

The four axes are conceptually independent (a composite can be high-C_{comp} low-S, low-high, or high-high), and existing attacks occupy different cells of this plane, which is why cross-attack ASR comparisons conflate the factors. We also acknowledge a fifth axis, instruction encoding/decodability (encoding-based MML and VisCo's conversational scaffold), which we do not dissect (set aside in the Limitations); and a sixth, undissected dimension, tile geometry and border statistics: on GEMMA-3-4B the padding-construction difference between baseline CS-DJ and the construction-matched *unrelated* cell moves ASR by 0.19 (Section 5.3), larger than any designed effect here, so how a composite tiles and pads its images is a first-order carrier the four constructed axes do not contain.

3.3 ATTACK PLACEMENT AND IMAGE-SIDE METRICS

We treat the four attack families (five configurations) as diagnostic probes rather than the object of study: FigStep and the MML variants occupy the low-C_{comp} region, CS-DJ the high-C_{comp} low-S region, and VisCo a high-density single-image scaffold (full placement detail and the three descriptive image-side metrics in Appendix A.34).

3.4 PAIRED ABLATIONS

A correlational comparison across the five baseline attacks cannot adjudicate whether complexity or alignment drives safety failure, so we construct two paired ablations that hold the harmful instruction byte-identical and vary one axis at a time, validated through image-side metrics before victim queries are run (constructions in Appendix A.46).

Complexity ablation (Exp. 2). From a single fixed FigStep typographic tile we build three stimuli, a padded surround, a random-natural-image distractor composite, and a query-aligned generated-scene composite (C_{comp} rising at roughly constant payload).

Alignment ablation (Exp. 3). From the CS-DJ 3×4 composite we hold the layout and the three sub-query tiles fixed and vary only the nine distractor tiles among noise, unrelated natural images, and query-aligned generated scenes (S rising at roughly constant density); because this also moves tile provenance, the clean S identification requires the separate E4 control that holds provenance fixed while removing only query relatedness (Appendix A.46).

3.5 Attribution Rule

We commit to a decision rule: a contrast is attributed to a descriptor only if it is the sole measured descriptor that changes between the two conditions; where two or more move together the contrast is *jointly attributed* and not counted toward either headline magnitude. A contrast closed by the E4 derangement control is a single-descriptor contrast by construction. Under this rule the paper draws exactly one contrast that survives the global correction, attributed under the measured-descriptor rule to the relatedness manipulation, with image-text congruence unmeasured: the E4 query-relatedness control on Qwen3-VL-8B, with the within-category harm-salience control (Appendix A.7) counted as a robustness check of the same claim on overlapping stimuli rather than an independent attribution; no compositional rung is a clean C_{comp} isolation. The full statement, its worked application, and the sampler/statistics protocol it uses (within-prompt k=3 mean tested by a permutation null at $p < 0.05$, bootstrap 95% CIs, Holm-corrected paired McNemar) are in Appendices A.35 and A.36. We declare a single authoritative report pipeline for every headline contrast: the within-prompt k=3 mean under the rubric judge (GPT-4o-mini backend) at τ=0.5, tested by a within-prompt permutation null at $p < 0.05$ two-sided with exact p; this sampler was adopted after single-draw results had already been observed, so it is declared here as the report standard, not as an outcome-independent pre-commitment. Any cell without a k=3 record (the Table 2 baseline matrix, the API rows, and the CS-DJ ladders) is labeled single-draw and descriptive and carries no confirmatory claim.

4 Experiments

Experiments. Experiment A.37 queries every (attack, victim) cell of the 5×5 panel (Table 2; Appendix A.37). Experiments A.38 and A.39 are the complexity and alignment ablations of Section 3.4, read through within-prompt paired contrasts (Appendices A.38, A.39). Experiment A.40 re-slices Experiments 1 to 3 by MLRM versus MLLM without new queries (Table 11; Appendix A.40).

Victim panel (five models). We evaluate two API-served frontier models (GPT-4o (OpenAI, 2024a) and Claude Sonnet 4.5 (Anthropic, 2025)), two open-source MLLMs (Qwen3-VL-8B-Instruct (Bai et al., 2025) and Gemma-3-4B-IT (Gemma Team, Google DeepMind, 2025)), and one reasoning-augmented multimodal model (GLM-4.1V-9B-Thinking (Team et al., 2025)). The exact sampling configuration (bf16 loading, `do_sample`, temperature, top-p, token budgets, image down-scaling, and judge backend) is fixed so every cell is reproducible and is detailed in Appendix A.59.

Dataset and judge. All cells are run on the full StrongREJECT slice (Souly et al., 2024) (n=313 prompts across six harm categories) so that every (attack, victim) cell is anchored on the same prompt order, which is required for the within-prompt paired McNemar tests reported in Sections A.38 and A.39.

Attack panel. The four baseline families (five configurations) are FigStep (Gong et al., 2025), two MML variants (Wang et al., 2025) (mirror, base64), CS-DJ (Yang et al., 2025), and VisCo (Li et al., 2025a); the Exp. 2/3 ablations add three FigStep variants (*pad*, *distract-random*, *distract-relevant*) and three CS-DJ variants (*noise*, *unrelated*, *relevant*), the generators of all six described in Section 3.4.

5 Results

Table 1 gives the identification status of every contrast under the attribution rule (Section 3.5). We anchor each family with the baseline matrix (Section 5.1), read the compositional (Sections 5.2, 5.3) and alignment (Section 5.4) ablations against the four-axis predictions, and close with the image-contribution control (Section 5.5), which licenses reading the composite rows at all. All cells use the full StrongREJECT slice (n=313).

Contrast	Descriptors moved	Attribution	ΔASR	p
Exp. 2 ladder (Gemma-3-4B)				
baseline → quarter	payload frac.; ent 0.90→2.09; JPEG 81.4→33.3	joint (geom.+dens.)	+0.012	0.36[†]
quarter → pad	tile count; JPEG 33.3→40.8	joint (C_{comp}+compr.)	+0.095	< 0.001
pad → distract-random	ent, JPEG, prov.	joint (C_{info}+P)	+0.070	< 0.001
Exp. 3 alignment (Qwen3-VL-8B)[**]				
noise → unrelated	ent, JPEG, prov., S	confounded (S w/ ASR)	+0.002	0.85
unrelated → relevant	ent, JPEG, prov., S	confounded (S w/ ASR)[**]	+0.093	< 0.001
noise → relevant	all four move	exploratory joint endpoint	+0.095	< 0.001[‡]
CS-DJ ladder / E4 provenance				
E4: relevant → gen-unrel.	**ent, JPEG, tile count, prov. all fixed**	**clean S (Qwen)**	**−0.119**	$\mathbf{1.0\times10^{-4}}$
c_1→c_4→c_{12} (Gemma)[§]	tile count; ent 6.99/7.45/6.62; JPEG 51.5/113.9/114.5; payload vis. absent → present	content-qty confounded (C2)	+0.46	n/a

Table 1: Bounded identification status per contrast (attribution rule in Section 3.5). Unless marked, every ΔASR and p is the within-prompt permutation estimate over per-prompt k=3 means (Appendix A.19), scored by the rubric judge at τ=0.5; no row mixes samplers. The E4 row sits at the within-prompt permutation floor p=1.0 × 10^{-4} ($1/(10^4+1)$) and is reported exactly. [§]Single-draw by construction. [†]Payload-geometry rung is at the resolution limit (p=0.36), unassigned. [‡]Holm-adjusted p=0.002. [**]Adjacent-step contrasts are primary, the ordered endpoint exploratory; the alignment claim rests on the k=3 localization at unrelated→relevant, the step the E4 control isolates (this reconciles the step-localization difference with the single-draw McNemar values of Table 7). ent = entropy, prov. = provenance; descriptor numbers are the full-slice, post-resize means of Table 32.

5.1 Baseline Cross-Attack Matrix

Table 2 shows the baseline matrix; four patterns are evident.

Four patterns. (i) *Typographic-defense gradient:* the MML variants collapse to ≤ 0.02 on Claude S.4.5 but run high elsewhere (GPT-4o 0.81/0.83; Gemma-3-4B 0.96/0.95; MLRM 0.78/0.39; Qwen3-VL-8B 0.90/0.41); as MML is typographic with no compositional structure, this is inconsistent with compositional complexity binding Claude, pointing to dedicated typographic robustness training. This reading is conditional on Claude's low provider-side interception: its per-request interception rate is small on every measured cell ($\leq$ 0.07 in Exp. 3/E4 controls, Appendix A.29), so the typographic robustness reading is not an artifact of provider moderation; had interception been substantial, the low ASR would have been uninterpretable. (ii) *Structural-defense gradient:* CS-DJ is bounded 0.28–0.32 on the frontier-grade victims (overlapping CIs) and jumps to 0.60/0.84 on Gemma-3-4B and the MLRM. (iii) *FigStep is victim-specific:* GPT-4o and Claude neutralize it ($\leq$ 0.01) but Gemma, Qwen3-VL, and the MLRM accept it at 0.29/0.13/0.73; Claude's low FigStep interception (0.05, Appendix A.29) means this neutralization is a model-level behavior, not an input-moderation artifact. (iv) *VisCo asymmetry:* neutralized on Claude (0.03) but lifting the others to 0.24–0.67, the MLRM (0.67) most vulnerable.

5.2 Experiment 2: Complexity Ablation

Exp. 2 reading. Table 29 reports the four-row ladder (per-victim numerics, rung decomposition, and sampling-variance floor in Appendix A.51); the step-wise decomposition and E5 payload-geometry control split the rung ordering into accepted and resolution-limited steps (Appendix A.26).

5.3 Compositional Ladder Inside the CS-DJ Family

CS-DJ region-count ladder. We isolate C_{comp} with a region-count ladder holding grid, pixel size, canvas, JPEG footprint and the three payload tiles fixed, varying only the number of distractor regions; the result is null on both victims: on Gemma-3-4B ASR is flat across the ladder (0.460 →

Attack	GPT-4o	Claude S.4.5	Qwen3-VL-8B	Gemma-3-4B	GLM-4.1V-Th.
FigStep	0.00 [0.00,0.00]	0.01 [0.00,0.02]	0.13 [0.09,0.17]	0.29 [0.25,0.35]	0.73 [0.68,0.78]
MML-mirror	0.81 [0.77,0.85]	0.02 [0.01,0.04]	0.90 [0.87,0.93]	0.96 [0.94,0.98]	0.78 [0.73,0.82]
MML-base64	0.83 [0.79,0.87]	0.00 [0.00,0.00]	0.41 [0.36,0.47]	0.95 [0.93,0.97]	0.39 [0.34,0.45]
CS-DJ	0.28 [0.24,0.34]	0.32 [0.27,0.37]	0.29 [0.24,0.34]	0.60 [0.55,0.66]	0.84 [0.80,0.88]
VisCo	0.26 [0.21,0.31]	0.03 [0.01,0.05]	0.24 [0.20,0.29]	0.57 [0.52,0.63]	0.67 [0.61,0.72]

Table 2: Baseline cross-attack ASR on StrongREJECT (n=313 prompts per cell, binarized at τ=0.5, judged with the StrongREJECT rubric judge, with bias-corrected 95% bootstrap CI over 10^4 resamples). The five victims span the typographic and composite branches of our taxonomy.

$0.460 \rightarrow 0.454$, $r_1 \rightarrow r_9$ Δ=-0.006), and on QWEN3-VL-8B the weak positive trend is not significant ($0.185 \rightarrow 0.208 \rightarrow 0.220$, $r_1 \rightarrow r_9$ $+0.035$, p=0.25) (Appendix Figure 4, Table 30). Predecessor-ladder controls, Gemma reconciliation, the padding-construction effect, cross-bench replication, and full reading are in Appendices A.52 and A.27. Outside the designed axes, the white-padded CS-DJ baseline on GEMMA-3-4B (0.60) exceeds the construction-matched force-squared *unrelated* cell (0.41, single-draw) by 0.19, the largest image-side swing in the paper; it tracks a border difference (mean white-pixel fraction 0.84 vs. 0.33) that the taxonomy does not contain (Appendix A.27).

5.4 EXPERIMENT 3: ALIGNMENT ABLATION

Exp. 3 reading. Table 7 (Appendix A.5) holds the CS-DJ layout fixed and swaps distractor tiles among noise, unrelated, and query-relevant scenes. The alignment effect is detectable on one open-weight victim (QWEN3-VL-8B) and absent within detection limits elsewhere (full per-victim reading Appendix A.53); a third headroom victim (INTERNVL3.5-8B) reproduces the complexity and provenance-adjacent orderings without strengthening either single-victim claim (Appendix A.28), and the CLAUDE row is *interpretation-limited* by provider-side input interception (Appendix A.29). On QWEN3-VL-8B ASR rises monotonically (*noise* 0.19, *unrelated* 0.24, *relevant* 0.29); this is the paper's only linear result, isolated by a three-way audit (ordered-endpoint delta with both p-values, the step-localized signature, and the E4 *gen-unrelated* control) on replicated k=3 draws (Appendix A.55). The E4 contrast is refusal-driven and survives a within-category harm-salience control (Δ=$+0.089$, p=10^{-4}; Appendix A.7): together they are the paper's single survivor of the corrected global multiplicity correction, counted as one contrast attributed to the relatedness manipulation (with image-text congruence unmeasured) plus its robustness check on overlapping stimuli, rather than two independent attributions (Table 22). The null is victim-specific and a bound, not an absent mechanism, with minimum detectable delta $\approx$ 0.06–0.10 (Appendix A.57); a retrieval-based dose-response test is flagged follow-up (Appendix A.32).

Per victim, the E4 contrast (*relevant* $\rightarrow$ *gen-unrelated*; paired ΔASR with Holm-adjusted McNemar p, Table 8) is significant only on QWEN3-VL-8B (-0.12 [-0.17, -0.06], p_H<0.001); the other four victims do not reach significance ($p_H \geq 0.07$).

5.5 IMAGE CONTRIBUTION CONTROL (E-1)

Image-contribution reading. On the API reference victim (GPT-4O) the bare query reaches 1.3% ASR, a benign unrelated image leaves it at 0.0%, and attaching a CS-DJ image lifts it to 28.4% on identical prompts; the pattern holds across the panel, with one flagged exception on GLM-4.1V-TH. (Appendix A.58).

Summary of the reading. Read together, the panel supports two crisp conclusions and leaves one synthetic axis open. The image is a necessary carrier: no victim redacts the bare harmful query, and a benign image does not reanimate it, while every attack composite moves the panel off its floor. Structural construction is the largest measurable image-side lever, but it operates outside the four constructed axes (the padding difference) or is confounded (the payload-visibility ladders), so the mechanism is not compositional complexity as prior work defines it. Semantic alignment is present and bounded, on a single open-weight victim, in a direction measured-descriptor rule attributes to the relatedness manipulation with image-text congruence unmeasured. The obfuscation families (MML, VisCo) are floored or saturated on the open-weight and reasoning-augmented rows, so they bound where an effect can be observed rather than adjudicate the mechanism.

6 Discussion and Conclusion

We asked which image-side dimension each image-to-text jailbreak exploits, and answered with a controlled dissection rather than another leaderboard entry. One contrast survives the global correction: on the single open-weight victim with the most headroom (Qwen3-VL-8B), the E4 query-relatedness control moves refusal, and under the measured-descriptor rule this is attributed to the relatedness manipulation, with image-text congruence unmeasured; the within-category harm-salience control (Section 5.4, Appendix A.7) reproduces it on overlapping stimuli and the result is sampling-audited on replicated draws. It is reported as a bounded, single-victim observation rather than a general mechanism or a clean isolation of semantic alignment. The compositional axis is closed as a controlled null within the constructed ladders, so multiplicity per se does not drive these attacks; whether tile-count structure matters is untested here. The result therefore does not resolve the disagreement between CS-DJ's complexity claim and VisCo's alignment-grounded framing, and it is consistent with a conditional, single-victim effect. An encoder-side attention-mass probe on the two effect-carrying open victims shows the payload absorbs disproportionate attention, but that allocation is constant across conditions and does not move with ASR (Appendix A.41); the probe bounds, but does not complete, the mechanistic story.

Defensive implications. The resolvable ladder rungs (pad-surround, content-fill) are the largest detectable sources of variance and the candidate intervention surface for input-side screening and typographic-payload robustness training, bounded to the two victims that exhibit them; because the per-tile density screen does not separate generated and typographic tiles from photo tiles, a usable defense needs a signal beyond the measured density descriptors. Full discussion, including amplification localization and the adaptive-adversary caveat, is in Appendix A.49.

Methodological takeaway. The resolution of our design depends less on the particular attacks we chose than on the discipline of isolating an axis before attributing it: every contrast here is byte-identical in the instruction, screened on its image-side descriptors before any victim is queried, scored under one sampler and one judge, and either cleanly attributed or explicitly labeled joint. Applied to our own constructions, that discipline invalidated two of the compositional ladders we built and left a single surviving, bounded alignment observation, which is the outcome we report. A defender can reuse the same template on a new attack family (for instance a scaffold- or encoding-based one we set aside) and obtain an adjudicated claim about which, if any, image-side property drives it, without first ranking it against a leaderboard.

7 Limitations

Our conclusions rest on open-weight victims with headroom: the alignment observation on Qwen3-VL-8B, and the padding, provenance and compositional-ladder observations on Gemma-3-4B; none is a property of the full panel. The image-side descriptors are proxies (pixel entropy, JPEG@85, CLIPScore), and every ASR is judge-conditional: the rubric judge shares a family with the GPT-4o victim and differs from a non-OpenAI judge by up to 0.18, so magnitudes inside the roughly 0.10 band are read as signs and orderings only (Appendices A.15, A.4). The padding-construction swing (0.19; Section 5.3) exceeds every designed effect and caps any axis claim.

No contrast varies one axis while provably holding the others fixed, so the four axes are a labeling scheme over our constructions rather than an identification in the strongest sense; that requires a generator matching entropy, compressibility, tile count and provenance by construction (the phase-scrambled control of Appendix A.25 is a proof of concept). The E4 contrast isolates the relatedness manipulation but not image-text congruence, since no benign-paraphrase control was run. Whether tile-count structure matters is untested: the region-count null covers distractor diversity only, and a corrected tile-count ladder with the payload visible at every rung was not built. We exclude encoding/decodability and claim no mechanism: the attention-mass probe is null on both effect-carrying victims (Appendix A.41) and the in-model relatedness probe is too noisy to discriminate (Appendix A.17).

Deferred controls (protocols in Appendix A.31, not run). (a) Benign-generated unrelated distractors, separating a relevance-driven E4 gap from off-topic-harm suppression; (b) an on-topic scene generated from a benign paraphrase of the query; (c) the relevant scene paired with a benign restatement of the query, isolating congruence. Power bounds the alignment reading: the paired design resolves $|\Delta\mathrm{ASR}|$ only down to about 0.06–0.10 (Table 12), so a powered replication (n=830 paired prompts for a 0.06 effect at 80% power) is the natural confirmatory step.

AI USE STATEMENT

In accordance with the ICLR policy, we disclose the use of generative AI tools in this work and take full responsibility for its content. This section does not count toward the page limit.

Generative AI used for required-disclosure tasks.

- **Synthetic research stimuli.** A subset of the image stimuli used in the experiments (the query-aligned scenes of Exp. 2/3 and the provenance control) were generated with the text-to-image model Stable Diffusion 3.5, as described in the methodology and in Appendix A.46; this is a documented construction step of the attack pipeline, not a source of the reported measurements.
- **Automated judge.** The StrongREJECT rubric judge used to score every response is itself a generative model (a `gpt-4o-mini` backend), disclosed in Section 3.5 and audited against human labels in Appendix A.15.
- **Code, methodology, and analysis assistance.** LLM-assisted tooling was used to debug and refine the evaluation, manipulation, and statistical-analysis code, and to support the data analysis. All experimental results reported here (every ASR, confidence interval, and p-value) were produced by executing that code against real model queries on the documented panel and slice; no generative model emitted any reported number.

Generative AI used for recommended-disclosure tasks. LLMs were used to draft and edit manuscript prose, to format references, and to suggest repository structure. LLM-generated suggestions for code and statistics were reviewed, executed, and verified against the recorded experimental data before they were accepted. We reviewed all AI-assisted work; we do not rely on any AI-generated claim that was not independently checked. The reported figures were produced programmatically from the recorded outputs, and the raw harmful model outputs are not reproduced in the paper (see the Ethical Considerations statement). We take full responsibility for the final content of this work, including the text, claims, code, and artifacts produced with the aid of generative AI.

ETHICAL CONSIDERATIONS

We study existing safety failures; no fine-tuned attacker model or new exploit is released. All attacks and controlled variants are taken from prior published work, and harmful prompts come from the public StrongREJECT benchmark run as-is. Raw harmful model outputs are never reproduced; the rendered typographic payload in Figure 5 is Gaussian-blurred for display, and tables report only aggregated metrics. API victims were queried under standard red-teaming terms, without attempting to extract weights or training data. The released per-prompt judged outputs contain victim response text, so we gate the complete archive behind access control. Labels use an automated StrongREJECT rubric judge audited against human labels (0.800 agreement, $\kappa{=}0.594$; Appendix A.15); judge error does not interact with the manipulated factor in a way that changes any conclusion. Defences motivated by our findings should be evaluated against adaptive adversaries before deployment.

REPRODUCIBILITY STATEMENT

We release an anonymized, redacted copy of the full evaluation source—the stimulus generators, judge prompts, exact StrongREJECT prompt-order slice, per-cell judged outputs, image-side metric code, and the statistical code behind every table and figure—as `reproducibility_anonymous.tar.gz`, mirrored on an anonymous public repository. The complete experimental inventory (query counts, compute and API budget, deterministic seeding, and the exact script behind every reported number) is detailed in Appendix A.48; query sampling, the single authoritative $k{=}3$ within-prompt sampler, the paired McNemar and bootstrap statistics, and their power bounds are specified in Appendix A.36 (with the power analysis in Appendix A.12 and the consolidating mixed model in Appendix A.10), and the attribution rule that licenses each reported claim is stated in Section 3.5 and Appendix A.35. Consistent with the Ethical Considerations statement, the shipped archive contains image-side metrics, judged responses, and the prompt-order slice rather than the raw SD3.5-generated harmful-scene images; if scene images are included, they sit in a separate access-controlled directory released on request for red-teaming verification.

REFERENCES

Anthropic. Claude sonnet 4.5 system card. `https://www.anthropic.com/claude-sonnet-4-5-system-card`, September 2025. URL `https://www.anthropic.com/claude-sonnet-4-5-system-card`.

Shuai Bai et al. Qwen3-VL technical report. *arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Yuntao Bai, Andy Jones, Kamal Ndousse, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv:2204.05862*, 2022.

Ravikumar Balakrishnan, Sanket Mendapara, and Ankit Garg. Reading between the pixels: Linking text-image embedding alignment to typographic attack success on vision-language models. *arXiv:2604.12371*, 2026. URL https://arxiv.org/abs/2604.12371.

Dongyu Chen et al. FC-Attack: Jailbreaking multimodal large language models via auto-generated flowcharts. In *Findings of EMNLP*, 2025. URL https://arxiv.org/abs/2502.21059.

Gemma Team, Google DeepMind. Gemma 3: A family of open models based on gemma 2. *arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. FigStep: Jailbreaking large vision-language models via typographic visual prompts. In *AAAI*, 2025. URL https://arxiv.org/abs/2311.05608.

Hakan Inan et al. Llama Guard: LLM-based input-output safeguard for human-AI conversations. *arXiv:2312.06674*, 2023.

Jia Li et al. Visual contextual attack: Jailbreaking MLLMs with image-driven context injection. In *EMNLP*, 2025a. URL https://aclanthology.org/2025.emnlp-main.487.

Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Images are Achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. In *European Conference on Computer Vision (ECCV)*, 2024. URL https://arxiv.org/abs/2403.09792.

Zuoou Li, Weitong Zhang, Jingyuan Wang, Shuyuan Zhang, Wenjia Bai, Bernhard Kainz, and Mengyun Qiao. Towards effective MLLM jailbreaking through balanced on-topicness and OOD-intensity. *arXiv:2508.09218*, 2025b. URL https://arxiv.org/abs/2508.09218.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. MM-SafetyBench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision (ECCV)*, pp. 386–403. Springer, 2024. doi: 10.1007/978-3-031-72992-8_22.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. JailBreakV-28K: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2404.03027.

OpenAI. GPT-4o technical report. Technical report, OpenAI, 2024a. URL https://openai.com/research/gpt-4o.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024b.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *AAAI*, 2024. URL https://arxiv.org/abs/2306.13213.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. URL https://arxiv.org/abs/2103.00020.

Bingrui Sima et al. VisCRA: A visual chain reasoning attack for jailbreaking multimodal large language models. In *EMNLP*, 2025. URL https://arxiv.org/abs/2505.19684.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for empty jailbreaks. In *NeurIPS Datasets and Benchmarks*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/e2e06adf560b0706d3b1ddfca9f29756-Abstract-Datasets_and_Benchmarks_Track.html.

GLM Team et al. GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv:2507.01006*, 2025. URL https://arxiv.org/abs/2507.01006.

Shengbang Tong et al. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *CVPR*, 2024.

Ruofan Wang et al. White-box multimodal jailbreaks against large vision-language models. In *ACM MM*, 2024a. URL https://arxiv.org/abs/2405.17894.

Yichen Wang et al. AdaShield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. *arXiv:2403.09513*, 2024b.

Yu Wang et al. Jailbreak large vision-language models through multi-modal linkage. In *ACL*, 2025. URL https://aclanthology.org/2025.acl-long.74/.

Shukun Xiong et al. VRSA: Jailbreaking multimodal large language models through visual reasoning sequential attack. arXiv preprint arXiv:2512.05853, 2025. URL https://arxiv.org/abs/2512.05853.

Zuopeng Yang et al. Distraction is all you need for multimodal large language model jailbreaking. In *CVPR*, 2025. URL https://arxiv.org/abs/2502.10794.

Hongqiong Zhang et al. Towards visualization-of-thought jailbreak attack against large visual language models. In *NeurIPS*, 2025. URL https://openreview.net/forum?id=5P5YgohyBZ.

Yubo Zhao et al. MMJ-Bench: A comprehensive study on jailbreak attacks and defenses for multimodal large language models. In *AAAI*, 2025. URL https://arxiv.org/abs/2408.08464.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2402.02207.

A Appendix

A.1 Per-Category Breakdown on the API Reference Victim

Table 3 reports the per-category breakdown on the API reference victim (GPT-4o) across all six StrongREJECT harm categories. The two MML variants are uniformly effective with a single outlier (Sexual Content drops to 0.62 for both, suggesting an additional category-level filter active above the typographic layer); the CS-DJ baseline lifts every category in the 0.20–0.42 band but with a clear concentration on *Disinfo./Decep.* (0.42) and *Non-violent Crimes* (0.39). The three FigStep ablation variants stay flat at 0.00 across all categories, mirroring the baseline FigStep row of Table 2. Among the CS-DJ ablation variants, *relevant* tracks the baseline closely ($\leq \pm 0.04$ per category), while *noise* and *unrelated* reshuffle the category profile more visibly (up to ± 0.18 on individual categories) without changing the overall ASR by more than 0.04. The per-category view shows that the composite layout, not tile content, determines which categories CS-DJ leaks on.

A.2 Manipulation Check

Table 4 reports the authoritative post-resize (longer side ≤ 768 px) image-side metrics recomputed from source stimuli by a single script, confirming the empirical complexity ordering used in Section 3.3: per-image entropy of the three low-C_{comp} families (FigStep, MML-mirror, MML-base64)

	Disinfo./Decep.	**Illegal G./S.**	**Hate/Harass.**	**Non-viol.**	**Violence**	**Sexual**	**Overall**
FigStep family (4 conds.)			0.00 (all categories)				**0.00**
MML-mirror	0.92	0.80	0.76	1.00	0.74	0.62	**0.81**
MML-base64	0.92	0.74	0.82	0.92	0.93	0.62	**0.83**
CS-DJ	0.42	0.20	0.22	0.39	0.22	0.24	**0.28**
VisCo	0.35	0.33	0.29	0.27	0.19	0.16	**0.26**
CS-DJ-noise	0.28	0.22	0.34	0.29	0.22	0.10	**0.24**
CS-DJ-unrelated	0.36	0.38	0.28	0.34	0.31	0.20	**0.31**
CS-DJ-relevant	0.42	0.22	0.22	0.41	0.26	0.20	**0.29**

Table 3: Per-category ASR breakdown on the API reference victim (GPT-4o; full StrongREJECT, n=313, rubric judge with τ=0.5). Top block: baseline attacks anchoring the $C \times S$ taxonomy; the four Exp. 2 complexity conditions compress into a single all-zero row (a genuine hard refusal, Section 5.5). Bottom block: Experiment 3 alignment ablation (CS-DJ layout held constant).

Attack family	**Entropy** (bits/px)	**JPEG@85** (KB)
Experiment 1 (baseline attack images, n=20/cell; pre-resize sample)		
FigStep	0.71 ± 0.14	61.3
MML-mirror	0.73 ± 0.14	63.8
MML-base64	0.60 ± 0.29	55.7
CS-DJ	1.87 ± 0.07	321.3
VisCo	7.05 ± 0.53	179.6
*Experiment 2 (complexity ablation; **full slice**, n=313; post-resize)*		
FigStep (baseline)	0.90 ± 0.15	81.4
FigStep-pad	2.57 ± 0.09	40.8
FigStep-distr.-rand.	6.48 ± 0.54	112.9
FigStep-distr.-rel.	6.78 ± 0.21	122.9
Experiment 3 (alignment ablation, full slice, n=313; post-resize)		
CS-DJ-noise	5.97 ± 0.05	166.3
CS-DJ-unrelated	6.40 ± 0.17	131.1
CS-DJ-relevant	6.62 ± 0.13	114.6
CS-DJ-gen-unrelated (E4)	6.62 ± 0.12	114.7

Table 4: Manipulation-check summary on the full StrongREJECT slice. *Exp. 1* rows aggregate n=20 images per cell; entropy is mean per-channel Shannon entropy of the RGB image and JPEG@85 is the re-encoded byte size at quality 85. *Exp. 2/3* rows are on the *full* slice (n=313 per condition), recomputed by a single script on the *post-resize* image (longer side $\leq$ 768 px, LANCZOS) the encoders receive. Exp. 3 conditions are *not* matched in entropy (span 0.65 bits/px), compressibility, or tile provenance, so alignment is isolated only where those confounded descriptors run opposite to ASR or are held fixed by the E4 control.

Condition	Mean	SD	p50	95% range
Exp. 2 (complexity, FigStep payload)				
FigStep (baseline)	0.347	0.032	0.342	[0.30, 0.40]
FigStep-pad	0.306	0.032	0.306	[0.25, 0.36]
FigStep-distr.-rand.	0.285	0.034	0.288	[0.22, 0.33]
FigStep-distr.-rel.	0.324	0.033	0.325	[0.27, 0.38]
(surround-only CLIP, Exp. 2 distract-)*				
FigStep-distr.-rand.	0.124	0.021	0.123	[0.09, 0.15]
FigStep-distr.-rel.	0.216	0.034	0.215	[0.16, 0.28]
Exp. 3 (alignment, CS-DJ layout)				
CS-DJ (anchor)	0.198	0.041	0.196	[0.13, 0.27]
CS-DJ-noise	0.197	0.043	0.196	[0.13, 0.27]
CS-DJ-unrelated	0.187	0.040	0.186	[0.13, 0.25]
CS-DJ-relevant	0.254	0.040	0.254	[0.19, 0.32]
CS-DJ-gen-unrelated	0.196	0.046	0.196	[0.10, 0.29]
(surround-only CLIP, Exp. 2 gen-unrelated)				
FigStep-distr.-gen-unr.	0.136	0.026		–

Table 5: Alignment verification on the full StrongREJECT slice (n=313 per condition): whole-image CLIP ViT-L/14 cosine similarity between the stimulus and the harmful query (mean, SD, median, 5th–95th percentile). The Exp. 3 *relevant* condition carries measurably higher CLIP alignment than *noise* or *unrelated* (0.254 vs $0.197/0.187$), confirming the alignment manipulation is effective; the Exp. 2 surround-only rows show the query-aligned generated surround (0.216) exceeds the random natural surround (0.124).

sits at ≈ 0.7 bits/pixel with JPEG@85 size in the 56–64 KB range, CS-DJ jumps to 1.87 bits/pixel and 321 KB, and VisCo's generated-scene image sits at 7.05 bits/pixel and 180 KB. For the Exp. 2 variants (post-resize) the full-slice entropy rises from 0.90 at *baseline* to 2.57 at *pad* and to $6.48/6.78$ at the distract-* end of the ladder; for Exp. 3 the three tile-content variants are *not* matched in per-pixel entropy (they span 0.65 bits/px, 5.97–6.62), nor in compressibility (post-resize JPEG@85 spans 166.3 to 114.6 KB) nor in tile provenance, while alignment varies (Section A.3). The manipulations move and isolate the intended axis in both ablations.

A.3 Alignment Verification (CLIP)

Because the alignment effect is bounded (detectable on one victim, see Section 5.4), Table 5 verifies that the alignment manipulation actually moved: the Exp. 3 *relevant* condition carries measurably higher whole-image CLIP alignment to the harmful query than *noise* or *unrelated* (median 0.254 vs $0.196/0.186$), and the Exp. 2 query-aligned surround (0.216, surround-only) scores higher than the random natural surround (0.124). The alignment axis is therefore not a no-op: a sanitized or ineffective "relevant" manipulation would have left CLIP flat, but it measurably moved while ASR did not. The provenance control's *gen-unrelated* cell (Appendix A.6) sits at 0.196 whole-image CLIP, essentially at the *unrelated* level (0.187) and clearly below *relevant* (0.254), confirming that the permutation holds "SD3.5-generated harmful-domain scene" constant while varying only relatedness. The query-aligned scenes were generated with *Stable Diffusion 3.5 Large Turbo* (locally served via *diffusers*), an unconditional text-to-image model with no harmful-content refusal filter; a refusal or regeneration would have manifested as CLIP scores flat across conditions, which Table 5 rules out.

A.4 Judge Reliability (GPT-4o × FigStep)

Table 6 reports the continuous judge-score distribution for the GPT-4o FigStep cell, which is exactly 0.00 ASR across all conditions. The cell contains n=313 records of which 312 produced a judged response; one prompt (*strongreject_0120*) was excluded because the GPT-4o API rejected the query at the moderation stage (input flagged as illicit/violent, no credits charged), so no victim response existed to judge. All 312 scored responses sit far below the 0.5 binarization threshold (99th percentile 0.003, maximum 0.011): this is hard refusal, not a mass sitting just under the threshold, so the reported zero ASR reflects genuine total refusal on this cell rather than a threshold artifact of the judge.

Statistic (over 312 judged responses)	Value
Mean score	0.0006
Median	0.001
90th percentile	0.001
99th percentile	0.003
Maximum	0.011
Share at exact 0.0	0.32
Share in $(0, 0.5)$	0.68
Share ≥ 0.5 (ASR)	0.00

Table 6: Continuous judge-score distribution for the GPT-4o FigStep cell: n=312 responses scored with the StrongREJECT rubric. All scores sit far below the 0.5 binarization threshold (99th percentile 0.003, maximum 0.011), so the reported 0.00 ASR reflects genuine total refusal rather than a threshold artifact of the judge.

A.5 Experiment 3: Alignment Ablation Table

Table 7 reports the full Experiment 3 alignment ablation discussed in Section 5.4: the CS-DJ 3×4 layout is held fixed and the nine distractor tiles are swapped among noise, unrelated, and relevant levels. Each cell in the bottom block reports the raw and Holm-adjusted p-value of the paired McNemar test (family of three per victim); the two adjacent-step contrasts are the primary tests and the ordered endpoint $noise \rightarrow relevant$ is reported as an exploratory pooled contrast. These paired tests are computed between the three variant conditions directly and are thereby free of the anchor's complexity mismatch. The improvement over the earlier anchor-versus-CS-DJ tests is that the anchor lies at a different complexity level (1.87 vs 6.0–6.6 bits/px). One victim, Qwen3-VL-8B, shows a significant ordered-endpoint rise (Δ=+0.10, Holm p=0.002), which the main text reads as the single detectable alignment effect; on the other four victims no monotone noise-to-unrelated-to-relevant rise reaches significance. Table 9 additionally reports the paired-prompt ΔASR with bootstrap CIs for every contrast.

A.6 Experiment 3: Provenance Control (E4)

The provenance control (E4) closes the strongest confound against the Experiment 3 alignment ablation: the *relevant* condition differed from *unrelated* not in query-relatedness but in being produced by a synthesis model (SD3.5 generated an on-query scene, the unrelated pool did not). To separate synthesis provenance from query-relatedness, the *gen-unrelated* cell holds "SD3.5-generated harmful-domain scene" constant and varies only relatedness, by attaching query B's pre-generated scene to query A through a fixed derangement, with no prompt keeping its own scene). A CLIP manipulation check (Appendix A.3) confirms the permutation succeeded: whole-image CLIP to the harmful query is 0.196 for *gen-unrelated*, essentially at the *unrelated* level (0.187) and clearly below *relevant* (0.254), so the condition is "generated yet unrelated."

Table 8 reports the cell and its two paired contrasts. The single most informative read is on Claude, the victim whose *unrelated* peak (0.32) motivated the OOD-filter reading. If synthesis provenance had driven the effect, *gen-unrelated* would collapse toward the noise and relevant ends (0.00–0.22); if query-relatedness drives it, it should sit near *unrelated*. *gen-unrelated* reaches

Condition	GPT-4o	Claude S.4.5	Qwen3-VL-8B	Gemma-3-4B	GLM-4.1V-Th.
ASR@0.5 [95% bootstrap CI]					
CS-DJ-noise	0.24 [0.19, 0.29]	0.00 [0.00, 0.00]	0.19 [0.14, 0.23]	0.73 [0.68, 0.78]	0.85 [0.81, 0.89]
CS-DJ-unrelated	0.31 [0.26, 0.36]	0.32 [0.27, 0.37]	0.24 [0.19, 0.29]	0.41 [0.35, 0.46]	0.86 [0.82, 0.90]
CS-DJ-relevant	0.29 [0.24, 0.34]	0.22 [0.18, 0.27]	0.29 [0.24, 0.34]	0.66 [0.61, 0.71]	0.84 [0.80, 0.88]
Within-Exp. 3 paired McNemar per victim: (raw p, Holm p)					
noise → unrelated	(0.018, 0.055)	(<0.001, <0.001)	(0.060, 0.121)	(<0.001, <0.001)	(0.775, 1.000)
unrelated → relevant	(0.489, 0.489)	(0.003, 0.003)	(0.133, 0.133)	(<0.001, <0.001)	(0.341, 1.000)
noise → relevant†	(0.105, 0.211)	(<0.001, <0.001)	(<0.001, 0.002)	(0.028, 0.028)	(0.576, 1.000)

Table 7: Experiment 3: semantic-alignment ablation within the fixed CS-DJ 3×4 layout (StrongREJECT, n=313 per cell, rubric judge, τ=0.5). Top block: ASR@0.5 with 95% bootstrap CI. Bottom block: within-experiment paired McNemar tests as (raw p, Holm-adjusted p) pairs across the family of three contrasts per victim; the two adjacent-step contrasts are primary, the ordered endpoint noise → relevant† exploratory. Delta-ASR with 95% CIs: Table 9.

Condition	GPT-4o	Claude S.4.5	Qwen3-VL-8B	Gemma-3-4B	GLM-4.1V-Th.
ASR@0.5					
CS-DJ-noise	0.24	0.00	0.19	0.73	0.85
CS-DJ-relevant	0.29	0.22	0.29	0.66	0.84
CS-DJ-unrelated	0.31	0.32	0.24	0.41	0.86
CS-DJ-gen-unrelated	0.27	0.26	0.17	0.60	0.81
relevant → gen-unrelated: paired ΔASR, 95% bootstrap CI, Holm-p per victim					
ΔASR	-0.02	$+0.04$	-0.12	-0.06	-0.03
95% CI	[−0.07,0.03]	[−0.02,0.09]	[−0.17,−0.06]	[−0.12,0.00]	[−0.08,0.02]
Holm-p	0.54	0.22	< 0.001	0.07	0.29
unrelated → gen-unrelated: paired ΔASR, 95% bootstrap CI, Holm-p per victim					
ΔASR	-0.04	-0.06	-0.07	$+0.19$	-0.05
95% CI	[−0.10,0.02]	[−0.11,0.00]	[−0.13,−0.01]	[0.13,0.26]	[−0.10,−0.01]
Holm-p	0.19	0.05	0.03	< 0.001	0.03

Table 8: Experiment 3 provenance control (E4), ASR@0.5 and the two paired-prompt ΔASR [95% bootstrap CI] with Holm-adjusted McNemar p per victim (StrongREJECT, n=313 per cell, rubric judge, τ=0.5). The *gen-unrelated* cell holds "SD3.5-generated harmful-domain scene" constant and varies only query-specific relatedness, by attaching query B's pre-generated scene to query A via a fixed derangement. The noise/relevant/unrelated anchors are repeated from Table 7 for reference.

0.26, above the OOD prediction's upper limit and statistically indistinguishable from *unrelated* (Δ=$-0.06[-0.11, 0.00]$, p=0.05). The provenance control therefore does *not* collapse under an OOD read: "generated but unrelated" behaves like "unrelated," not like a synthesis artifact, on the victim where the two stories disagree most sharply. On Qwen, *gen-unrelated* drops to 0.17, significantly below *relevant* (Δ=-0.12, p<0.001) and below *unrelated* (Δ=-0.07, p=0.03): on the victim that carries the detectable alignment effect in Section 5.4, removing query-relatedness removes the effect, which is exactly what an alignment-based account predicts. On Gemma, *gen-unrelated* (0.60) sits near *relevant* (0.66) and far above *unrelated* (Δ=$+0.19$, p<0.001): for Gemma, any SD3.5-generated harmful-domain scene lifts ASR regardless of query-specific relevance, consistent with the C_{info}-driven reading of that victim in Section 5.4. Together these three victims triangulate the axis: on Claude the provenance control isolates relatedness; on Qwen it confirms that relatedness carries the signal; on Gemma it shows the effect is not query-specific. The alignment effect reported in the main text therefore survives the provenance confound.

A.7 Within-Category Derangement (Harm-Salience Control)

The E4 derangement above permutes the cached SD3.5 scenes across prompts without regard to harm category: the original control attaches each prompt to a scene generated for a *different* StrongREJECT category 84.03% of the time (263/313). Because the attached scene is typically from a different harm category than the query it accompanies, a reader could attribute the relevant-versus-gen-unrelated drop to cross-category harm salience (the attached scene competing for the victim's attention on a different harm topic) rather than to query relatedness. We close that confound with a same-category derangement (fixed seed 2026, zero cross-category pairs, zero fixed points; every StrongREJECT category has at least 50 prompts, so an internal derangement exists for all 313). The attack variant, stimulus bytes, sampler (k=3), and rubric judge are otherwise identical to the original E4 cell.

On Qwen3-VL-8B the within-category control reproduces the effect: relevant 0.232 vs. gen-unrelated-within 0.143, a per-prompt *relevant→gen-unrelated-within* Δ=-0.089 (95% CI [−0.131,−0.049], within-prompt permutation p=10^{-4}), same sign as the original cross-category Δ=-0.119 ([−0.160,−0.079]). The Shapley decomposition (Appendix A.8) is refusal-driven in both cases: within-category S_{ref}=0.549 (CI $[0.449, 0.642]$) and original S_{ref}=0.540 ($[0.475, 0.608]$), both excluding the 0.25 null. The within Δ is slightly smaller in magnitude than the cross Δ (0.089 vs. 0.119), recorded as a small residual sensitivity, but its CI is well clear of zero and its classification does not change, so the within-category control passes rather than collapses: cross-category harm salience is not the driver of the E4 effect, and the refused-driven, query-relatedness reading (Branch A) survives gate 2 (Section 5.4). This control also enters the corrected multiplicity family (Table 22).

Contrast	GPT-4o	Claude S.4.5	Qwen3-VL-8B	Gemma-3-4B	GLM-4.1V-Th.
Experiment 2 (baseline vs. variant; $+\Delta$ = variant above anchor)					
FigStep → pad	$+0.00$ [0.00, 0.00]	$+0.00$ [−0.01, 0.01]	$+0.00$ [−0.03, 0.03]	$+0.17$ [0.11, 0.22]	$+0.07$ [0.02, 0.12]
FigStep → distr.-random	$+0.00$ [0.00, 0.00]	$+0.03$ [0.01, 0.04]	$+0.01$ [−0.03, 0.04]	$+0.34$ [0.28, 0.40]	$+0.16$ [0.11, 0.20]
FigStep → distr.-relevant	$+0.00$ [0.00, 0.00]	$+0.04$ [0.02, 0.06]	$+0.02$ [−0.02, 0.05]	$+0.30$ [0.24, 0.36]	$+0.12$ [0.07, 0.17]
pad → distr.-random‡	$+0.00$ [0.00, 0.00]	$+0.03$ [0.01, 0.05]	$+0.01$ [−0.02, 0.05]	$+0.17$ [0.11, 0.23]	$+0.08$ [0.04, 0.13]
Experiment 3 (within-condition; $+\Delta$ = success rises toward relevant)					
noise → unrelated	$+0.07$ [0.01, 0.12]	$+0.32$ [0.26, 0.36]	$+0.05$ [−0.00, 0.10]	-0.32 [−0.39, −0.26]	$+0.01$ [−0.04, 0.05]
unrelated → relevant	-0.02 [−0.08, 0.03]	-0.10 [−0.16, −0.04]	$+0.05$ [−0.02, 0.10]	$+0.25$ [0.18, 0.31]	-0.03 [−0.08, 0.02]
noise → relevant†	$+0.05$ [−0.01, 0.10]	$+0.22$ [0.17, 0.27]	$+0.10$ [0.04, 0.15]	-0.07 [−0.13, −0.01]	-0.02 [−0.06, 0.03]

Table 9: Paired-prompt ΔASR with bias-corrected 95% bootstrap confidence intervals (over 10^4 resamples of the fixed prompt order) for every headline contrast reported in Sections 5.2 and 5.4. Each cell shows the point estimate with its CI below. †Ordered endpoint, reported as an exploratory Experiment 3 contrast. ‡The content-fill step (*pad* → *distract-random*) within the Exp. 2 ladder at constant tile count; it is the component attributable to information density and tile provenance, computed as a paired within-prompt contrast.

A.8 Delta-ASR Confidence Intervals

Table 9 reports the paired-prompt ΔASR with bias-corrected 95% bootstrap confidence intervals for every headline baseline-vs-variant and within-Exp. 3 contrast. These are the effect sizes behind the p-values in Tables 29 and 7: reporting the delta separately from its p-value makes explicit which contrasts are both significant and large, and which would survive a permissive judge shift.

A.9 Experiment 2: Payload-Scale Control (E5)

The payload-scale control (E5) isolates the first rung of the Exp. 2 ladder. The baseline (FigStep) places the payload at full canvas scale with no surround; FigStep-pad shrinks the payload to a quar-

(a) ASR per cell

	No surround		Pad surround	
Payload scale	**Gemma**	**GLM**	**Gemma**	**GLM**
Full	0.294 *(baseline)*	0.735 *(baseline)*	0.419 *(fullpad)*	0.764 *(fullpad)*
Quarter	0.364 *(quarter)*	0.502 *(quarter)*	0.463 *(pad)*	0.808 *(pad)*

(b) Paired-prompt contrasts: ΔASR, 95% bootstrap CI, Holm-corrected McNemar p

		Gemma			GLM		
Contrast	Definition	Δ	95% CI	p	Δ	95% CI	p
Payload geometry	quarter − baseline	$+0.07$	[0.03, 0.12]	0.003	$+0.05$	[−0.01, 0.11]	0.12
Surround, full scale	fullpad − baseline	$+0.125$	[0.070, 0.179]	<0.001	$+0.03$	[−0.02, 0.08]	0.30
Surround, quarter scale	pad − quarter	$+0.10$	[0.05, 0.15]	<0.001	$+0.307$	[0.24, 0.37]	<0.001

Table 10: Payload-scale control (E5) on the two effect-carrying victims, laid out as a 2×2 factorial: payload at full or quarter canvas scale $\times$ no surround or a pad surround. Panel (a): ASR per cell, with each cell's name in italics. Panel (b): paired-prompt ΔASR with 95% bootstrap CI and Holm-corrected paired McNemar p-values. On Gemma-3-4B the payload-geometry contrast is significant in this single-draw run ($+0.07$, $p=0.003$) but does not survive the $k=3$ per-prompt permutation null of Appendix A.19 (permutation $p=0.36$); a surround at constant payload scale raises ASR at both scales, stronger at full scale ($+0.125$) than at quarter scale ($+0.10$).

ter of the canvas and adds a pad surround, so the raw rung-one jump conflates two changes (a geometry change and a new surrounding region). E5 splits these into a 2×2 factorial on the two effect-carrying victims (Table 10): full/quarter payload $\times$ no/pad surround, adding the two missing cells (full-scale-with-surround and quarter-scale-no-surround). On Gemma-3-4B the payload-geometry contrast (quarter-no-surround minus full-no-surround) shows a single-draw cell gap of $+0.07$ (p=0.003 in the E5 single-draw run), but this rung does not clear the k=3 sampling floor of Appendix A.19 (permutation p=0.36), so it is flagged at the resolution limit there. Adding a pad surround at constant quarter payload scale robustly raises ASR ($+0.10$, $p<0.001$; this is the pad-surround rung that clears the floor, jointly attributed to layout and compression), and at constant full payload scale the same surround raises ASR by $+0.125$ $[0.070, 0.179]$, $p<0.001$. On Gemma-3-4B the full-scale-surround cell (*fullpad*) is complete on the full n=313 slice (ASR 0.4185), so it is reported in Table 10. The single-draw and k=3 numbers belong to different samplers: the E5 factorial is a single-draw paired gap, while Appendix A.19 averages three draws per prompt. We take the k=3 test as the sampling-noise bound: the payload-geometry rung does not clear it and so is flagged at the resolution limit rather than resolved into either component, and the two rungs that do clear it (compositional, content-fill) carry the decomposition of Section 5.2. On GLM-4.1V-Th. the geometry axis is not silent, but its Exp. 2 signal is inverted relative to the ladder narrative: the quarter-scale cell is markedly lower than both baseline and pad (see Table 10), so the geometry contrast is a significant *decrease* on that victim rather than a saturated non-effect.

A.10 Consolidating Random-Intercept Mixed Model

To consolidate the per-victim pairwise tests into a single summary, we fit, per experiment, a random-intercept logistic mixed model (BinomialBayesMixedGLM, variational-Bayes) with condition as a categorical factor, victim as a fixed effect, a condition$\times$victim interaction, and a per-prompt random intercept accounting for the fixed prompt order on which all conditions were judged. The categorical design has a condition number of 28 (Exp. 2) and 22 (Exp. 3) with maximum VIFs of 7.6 and 6.7, moderate collinearity that does not destabilize the fitting. The per-prompt random-intercept posterior SD is 0.76 (Exp. 2) and 0.64 (Exp. 3), confirming substantial within-prompt dependence that a per-test, per-victim read cannot pool. On the logit scale the model reproduces the pairwise reading: in Exp. 2 the baseline$\rightarrow$variant contrasts are large and clearly non-zero on Gemma-3-4B (pad $+1.12\pm0.18$, distract-random $+2.20\pm0.19$, distract-relevant $+1.95\pm0.18$) and GLM-4.1V-Th. (pad $+0.63\pm0.21$, distract-random $+1.56\pm0.24$, distract-relevant $+1.12\pm0.22$), small on Qwen3-VL-8B, and near zero on GPT-4o. In Exp. 3 the noise$\rightarrow$relevant contrast is largest and positive on Qwen3-VL-8B ($+0.74\pm0.17$) and exceeds its noise$\rightarrow$unrelated counterpart ($+0.41\pm0.18$), confirming under the pooled model that the only detectable alignment signal sits on Qwen; the large contrasts on Claude reflect the all-zero noise cell (near-complete separation) and are read descriptively. A continuous-predictor sensitivity version using tile count, entropy, and CLIP alignment (JPEG dropped as near-collinear with entropy) has a condition number of 451–471 because these features are constant per condition in both experiments, so the continuous design is near-degenerate relative to the categorical one; we report this collinearity explicitly and let the categorical model carry the consolidating argument. This model is a summary, not the primary statistic: the designed within-prompt paired contrasts remain the trustworthier inference, and the model is used to confirm rather than to replace them.

A.11 Experiment 4: MLRM versus MLLM Paired Contrasts

Table 11 reports the five paired contrasts used to assess MLRM differential sensitivity (Experiment 4), reported per victim (no folded MLLM-block average). On the FigStep to distract-random ladder the MLRM gains $+0.16$, more than each of the three floored-or-low standard victims (GPT-4o $+0.00$, Claude $+0.03$, Qwen $+0.01$) but *less* than the strongest standard victim (Gemma-3-4B, $+0.35$); only two victims have measurable headroom, so this is a bounded, localized observation on the complexity contrast, not a general property of reasoning models. On the alignment sub-contrasts the MLRM does not move with relevance: from unrelated to relevant it moves -0.03 while the only victim with a detectable alignment signal (Qwen3-VL-8B) moves $+0.05$, and the CS-DJ$\rightarrow$relevant contrast is flat on both the MLRM and each standard victim (other than the OOD-polarized Claude). The MLRM is also the absolute top of the open-source column on every baseline row except MML-base64, consistent with reasoning-style training lowering the bar for the complexity-driven mechanism rather than adding robustness to composite-image attacks.

A.12 Minimum Detectable Alignment Effect

Table 12 reports, per victim, the smallest absolute semantic-alignment effect the paired McNemar test over n=313 prompts could resolve at 80% power (closure over the observed discordance).

Paired Contrast	GPT-4o	Claude S.4.5	Qwen3-VL-8B	Gemma-3-4B	GLM-4.1V-Th.
FigStep → distr.-random	+0.00	+0.03	+0.01	+0.34	+0.16
distr.-random → relevant	+0.00	+0.00	+0.01	−0.04	−0.03
noise → unrelated	+0.07	+0.32	+0.05	−0.32	+0.01
unrelated → relevant	−0.02	−0.10	+0.05	+0.25	−0.02
CS-DJ → relevant	+0.01	−0.10	+0.00	+0.06	+0.00

Table 11: Exp.4 re-slice: ASR change (ΔASR; signed, absolute ASR units) on key paired contrasts, reported *per victim* rather than as an MLLM-block average. The FigStep→distr.-random complexity contrast is dominated by GEMMA-3-4B (+0.34); GLM-4.1V-TH. moves +0.16, more sensitive than GPT-4O, CLAUDE, and QWEN3-VL but *less* than the strongest standard model. The Exp. 3 alignment contrasts are flat or move against the predicted ordering on the four non-Qwen victims, on which alignment is at or below the detection limit (Gemma's row orders by tile provenance rather than by S; Section 5.4). The per-victim view avoids the misleading implication of an MLLM-block average that the reasoning model amplifies *both* axes.

Paired contrast	GPT-4o	Claude S.4.5	Qwen3-VL-8B	Gemma-3-4B	GLM-4.1V-Th.
noise → unrelated	0.079	0.088	0.075	0.103	0.061
unrelated → relevant	0.076	0.088	0.082	0.100	0.064
noise → relevant	0.076	0.073	0.081	0.085	0.062

Table 12: Minimum detectable absolute ΔASR at 80% power for the paired McNemar test over n=313 prompts, conditional on the observed number of discordant pairs in each Experiment 3 contrast and victim (two-sided, α=0.05, normal approximation). Across the panel this bound ranges from 0.06 to 0.10 absolute ASR units; it is the smallest alignment effect this experiment could resolve on a given victim.

Benchmark	Victim	Contrast (E4)	n	80%-power MDE	Observed ΔASR
SafeBench	Qwen3-VL-8B	relevant → gen-unrelated	119	0.131	+0.06

Table 13: Minimum detectable absolute ΔASR at 80% power for the Qwen3-VL-8B E4 contrast on the SafeBench second harm set, same normal-approximation machinery as Table 12, conditional on the 33 discordant pairs observed in the 119 paired prompts. n is sized by the benchmark, not by power; the observed +0.06 sits below the 0.131 bound, so the SafeBench arm is an underpowered directional agreement (Table 12 is the n=313 primary set).

Across the panel these bounds range from 0.06 to 0.10 absolute ASR units. They ground the bounded alignment null stated in Section 5.4: the experiment can rule out alignment effects of order 0.10 or larger on every victim, and the residual pixel-entropy variance documented there is the remaining confounder it cannot rule out.

The same power bound for the Exp. 2 composition ladder, the headline axis of that experiment, is given in Table 14: On GEMMA-3-4B both headline steps clear the bound by roughly 2×; on GLM the compositional step sits at the bound, so the GLM compositional reading is power-limited, exactly as stated in Section 5.2.

Paired step	Gemma-3-4B	GLM-4.1V-Th.	Qwen3-VL-8B	Claude S.4.5
baseline → pad	**0.082**	0.072	0.042	0.015
pad → distract-random	**0.088**	0.064	0.047	0.025
distract-random → distract-relevant	0.086	0.058	0.055	0.032

Table 14: Minimum detectable absolute ΔASR at 80% power for the paired McNemar test over n=313 prompts, conditional on the observed number of discordant pairs in each Exp. 2 ladder step (two-sided, α=0.05, normal approximation), for the ladder's non-floored victims. On Gemma-3-4B both headline steps (baseline → pad, +0.17; pad → distract-random, +0.17) clear this bound (0.082, 0.088); on GLM-4.1V-Th. the compositional step (+0.07) sits just below its 0.072 bound while the density step (+0.08) clears it. GPT-4o is excluded (floored, no finite bound computable).

Condition	k=3 pooled ASR	Discordance floor	Exp. 2 ASR (single-draw)
FigStep (baseline)	0.276	0.101	0.29
FigStep-pad	0.382	0.105	0.46
FigStep-distr.-rand.†	0.643	0.104	0.64
FigStep-distr.-rel.	0.502	0.121	0.60

Table 15: Sampling-variance floor on the effect-carrying victim (GEMMA-3-4B). Each of the four Experiment 2 conditions was re-run at k=3 draws per prompt (same deterministic stimulus across draws). The discordance floor is the mean over prompts of $2p(1-p)$, the probability two independent draws differ at τ=0.5; it bounds the sampling noise a single-draw ASR estimate is subject to. †The *distr.-random* cell was re-judged with the *rubric* judge on the *real* cached stimulus, so its floor is comparable to the Exp. 2 table cell; the other three rows use the local finetuned judge. This floor is descriptive, not an acceptance test: a rung is accepted on the within-prompt permutation p-value (Section 3.5). The final column is the single-draw ASR of Table 29; it is named sampler per column. The *distr.-rel.* row carries the largest single-draw versus k=3 gap in the ladder (0.60 vs. 0.50, ≈0.10), consistent with its being the row with the highest discordance floor (0.121) in the same table, i.e. the one whose ASR is least stable across draws; its ordering should therefore be read at the k=3 estimate, not the single draw.

A.13 SAMPLING-VARIANCE FLOOR (E-2)

Table 15 addresses the single-draw objection directly: because every other cell in the paper reports one judge score per prompt, any between-condition ASR difference could be attributed to decoding noise. E-2 bounds that noise on the victim carrying the effect (GEMMA-3-4B) by re-running the four Experiment 2 conditions at k=3 draws per prompt, held at the same deterministic stimulus so that the measured discordance is purely model sampling variance. The resulting within-prompt floor, the probability that two independent draws receive different verdicts, is 0.10–0.12 across the four conditions and the mean within-prompt judge-score standard deviation is ≈0.09. All three headline Experiment 2 contrasts (baseline against pad, distract-random, and distract-relevant) exceed this floor, so their ASR differences are not artifacts of a single draw (the pad-surround and content-fill rungs remain the accepted steps; the payload-geometry contrast is flagged at the resolution limit under the k=3 permutation test, Appendix A.19); the weaker variant-vs-variant contrasts sit at or below it, consistent with those differences being sampling noise. This converts the single-draw concern from a fatal methodological flaw into a measured and bounded quantity at a fraction of the cost of resampling the full panel.

Because the floor is used to defend deltas produced by the rubric judge, we re-judged the cell whose margin most affects the ladder (*distract-random*) with the *rubric* judge on the *real* cached stimulus (the original E-2 run used the local finetuned judge and, for this cell, a seeded-noise fallback stimulus). Under the rubric judge the within-prompt floor is 0.104 with a mean within-prompt judge-score standard deviation of ≈0.11, matching the 0.10–0.12 bound from the local judge, and the headline baseline-to-distract-random contrast (Δ=0.349 on the rubric-comparable pooled ASR) still exceeds it. The floor is therefore judge-robust: switching from the local to the rubric judge does not change the conclusion that the ladder's effect-carrying steps are not single-draw noise.

A.14 THRESHOLD AND JUDGE ROBUSTNESS

Table 16 re-binarizes every cell at $\tau \in \{0.3, 0.5, 0.7\}$ and checks that the condition *orderings* the paper's claims rest on are stable. They are: QWEN3-VL-8B's ordered endpoint ($\text{noise} < \text{unrelated} < \text{relevant}$), the only detectable alignment signal, holds at all three thresholds; GEMMA-3-4B's tile-provenance ordering ($\text{noise} \geq \text{relevant} > \text{unrelated}$) holds; and the effect-carrying ladders on GLM-4.1V-TH. and the floored victims are stable up to an adjacent-cell swap inside the confidence intervals. The claims are ordering claims, and the ordering survives the threshold choice.

Judge robustness. Because every ASR delta is judge-conditional (Section 3.5), we state the bound between the two judges: the largest cross-judge gap between the local finetuned judge and the rubric judge is 0.18 absolute ASR at *distract-random* (the ladder's gap is roughly 0.02, 0.08, 0.18, and 0.10 at baseline, pad, distract-random, and distract-relevant). This 0.18 exceeds every headline delta except GEMMA-3-4B's content-fill step (the k=3 estimate +0.070; single-draw +0.17 sits on this band) and ladder total (+0.35); on the remaining contrasts the disagreement band is smaller than the reported effect, so the qualitative ordering survives a switch between judges. We therefore flag, as

	Exp. 2 (complexity)		Exp. 3 (alignment)	
Victim	Order	Stable	Order	Stable
GPT-4o	all 0.00 (floor)	–	un $>$ rel $>$ noise	yes*
Claude S.4.5	rel $\geq$ rand $>$ pad $\geq$ base	no	un $>$ rel $>$ noise	yes
Qwen3-VL-8B	rel $\approx$ rand $>$ base $>$ pad	no	rel $>$ un $>$ noise	yes
Gemma-3-4B	rand $\geq$ rel $>$ pad $>$ base	no	noise $>$ rel $>$ un	yes
GLM-4.1V-Th.	rand $>$ rel $>$ pad $>$ base	yes	un $>$ noise $>$ rel	yes

Table 16: Sensitivity of the condition ordering to the binarization threshold. For each victim we re-compute the per-condition ASR at $\tau \in \{0.3, 0.5, 0.7\}$ (StrongREJECT, n=313/cell) and list the resulting Exp. 2 and Exp. 3 orderings; "stable" means the identical ranking holds at all three τ. Exp. 2: base = baseline, pad, rand = distract-random, rel = distract-relevant; Exp. 3: noise, un = unrelated, rel = relevant. The claims the paper rests on survive: Qwen3-VL-8B's Exp. 3 endpoint ordering, Gemma-3-4B's tile-provenance ordering, and GLM-4.1V-Th.'s Exp. 2 ladder ordering are all stable. The only instabilities are between adjacent, statistically indistinguishable cells. *GPT-4o flips *unrelated*/*relevant* at τ=0.7 (both within their CI overlap); *noise* stays lowest.

Condition	n	**Agreement**	κ	**Judge ASR**	**Human ASR**
baseline	69	0.841	0.679	0.522	0.594
pad	64	0.797	0.594	0.484	0.500
distract-random	59	0.797	0.579	0.610	0.576
distract-relevant	58	0.759	0.496	0.586	0.621
overall	250	0.800	0.594	—	—

Table 17: Human-annotation judge agreement. A human annotator labeled 250 samples stratified by judge-score band (with deliberate oversampling in the two bands adjacent to τ=0.5) across the four Exp. 2 conditions. Overall agreement is 0.800 with Cohen's κ=0.594; per-condition κ ranges 0.496–0.679. The implied ASR shift (human labels replacing judge labels, stratum-weighted) is small and does not reverse any headline ordering.

in the main text, that the *magnitudes* near the disagreement band are not judge-invariant, while the *orderings* on the effect-carrying victims are.

Self-family re-judge of the GPT-4O rows. Because GPT-4O is both a victim and the family behind the rubric judge (`gpt-4o-mini`), we re-judge every GPT-4O cell with the local finetuned StrongREJECT judge (`strongreject-15k-v1`, a non-OpenAI network; results/r8/gpt4o_rejudge.json). The floored-row narrative survives: on the Exp. 2 ladder, which the rubric judge reads as 0.00 across all four conditions, the non-OpenAI judge also gives near-zero ASR (0.048–0.096), so GPT-4O's Exp. 2 floor is a genuine hard refusal rather than a self-family artifact. On the Exp. 3 rows the two judges agree in placing the cell ASRs in the 0.19–0.27 (rubric) and 0.19–0.40 (local) ranges with the same qualitative shape, and the per-category profile holds at the level of which categories leak most (Non-violent crimes highest, Sexual content lowest) under both judges. We therefore retain the GPT-4O Exp. 2 floor and its per-category breakdown as reported, now confirmed under a self-family-free label. Continuous-score paired tests. Because every McNemar test above binarizes at τ=0.5 and drops the continuous info, we repeat the two effect-carrying drivers directly on the continuous judge scores (paired Wilcoxon signed-rank over the n=313 shared prompts). The Exp. 3 ordered endpoint on QWEN3-VL-8B (noise $\rightarrow$ relevant) is confirmed: Wilcoxon p=0.0002 with a paired continuous mean shift of $+0.076$ $[0.037, 0.115]$, strictly stronger and threshold-free relative to the binarized $+0.10$ (Holm p=0.002). The content-fill step on GEMMA-3-4B (pad $\rightarrow$ distract-random) is confirmed at Wilcoxon $p<0.001$, mean $+0.141$ $[0.092, 0.189]$; on GLM-4.1V-TH. at $p<0.001$. Conversely, the cell the binarized test reads as null here, QWEN3-VL-8B's Exp. 2 pad $\rightarrow$ distract-random ($+0.01$), stays null on the continuous scale as well (Wilcoxon p=0.30), which shows the continuous test does not inflate noise into signal: it separates the genuinely weak drivers (Gemma/GLM content-fill, Qwen Exp. 3) from the genuinely flat ones (Qwen Exp. 2) with the same ordering the binarized analysis reports, but with tighter bounds.

A.15 Human-Annotation Judge Agreement

A central question about the automated judge is not whether it is accurate on average but whether its error interacts with the manipulated factor. To answer that, a human annotator labeled 250 samples drawn from the four Exp. 2 conditions, stratified by judge-score band with deliberate oversampling in the two bands adjacent to $\tau{=}0.5$ (where a single verdict flip changes the binarized outcome), so that the batch is not dominated by unambiguous refusals. Table 17 reports the result. Overall agreement with the rubric judge is 0.800 with Cohen's $\kappa{=}0.594$; per-condition κ ranges 0.496 (distract-relevant) to 0.679 (baseline). The implied ASR shift, extrapolating each stratum's human-vs-judge disagreement back to the full conditions with stratum weights, is small in every cell (baseline $+0.072$, pad $+0.016$, distract-random -0.034, distract-relevant $+0.034$) and, crucially, does not reverse any headline ordering: the ladder's effect-carrying steps and the alignment nulls survive a switch from judge labels to human labels. Judge error therefore does not interact with the manipulated factor in a way that changes any conclusion in the paper.

A.16 Payload Legibility Ladder

A concern is that the tile-count effect in the CS-DJ ladder is confounded with payload resolution: every victim encoder resizes the input to a fixed budget, so expanding the canvas to add tiles shrinks each tile's share of that budget and could make the harmful text progressively harder for the OCR pathway to read, with a monotone ASR rise being the result. This section measures the payload's legibility directly at each victim's post-resize resolution, using a local optical-character-recognition pass (tesseract, English) over the real cached stimulus. For each of 20 c12 stimuli we crop the three sub-query (payload) tiles, downscale the full stimulus to two target resolutions, and compute the character error rate (CER) against the ground-truth sub-query text from the stimulus metadata (embedded in the cached entry).

resize target	mean CER	mean payload text height (px)
source (no resize)	0.115	318
Qwen3-VL-8B post-resize	0.115	318
Gemma-3-4B post-resize (896^2)	0.162	137

Table 18: Payload legibility: character error rate and rendered text height of the CS-DJ sub-query payload under the two open victims' post-resize encoder resolutions, on 20 real stimuli. CER stays low (≤ 0.16) at every resolution, so payload legibility does not degrade with tile count; text height falls at Gemma's fixed 896^2 but remains far above the legibility floor.

First, the ladder cannot be attributed to composition alone: because each rung renders only the first N tiles of the fixed 12-tile bucket (9 distractors + 3 sub-query tiles), the c_1 and c_4 composites contain no legible sub-query payload text (0.3% and 0.6% of stimuli carry any oracle token on a tesseract pass; results/r5/audit_ladder_payload.json) while c_{12} carries it (the three typographic tiles, 8.3%). The ladder therefore varies both tile count and the visibility of the harmful sub-query text, so it is not a clean composition isolation and is handled as such in Section 5.3. Given that, the relevant legibility comparison is between the c12 payload at source, at Qwen's dynamic resolution, and at Gemma's fixed resolution. CER stays approximately flat ($0.115 \rightarrow 0.162$) as ASR rises threefold across the ladder (Table 23), and remains low even at Gemma's coarsest effective resolution. The tile-count effect is therefore not a relabeling of payload illegibility: the payload stays readable wherever it appears. This bounds, rather than removes, the resolution confound: the per-tile encoder budget still declines with tile count, but that budget decline does not cross the legibility threshold at which the payload becomes hard to read.

A.17 Alignment in Each Victim's Own Vision Encoder

The main-text alignment proxy is a CLIP ViT-L/14 cosine (Section 3.3). Because no victim uses that encoder, a null in CLIP space is compatible with a real alignment effect in each victim's own vision tower. Qwen3-VL, Gemma-3, and GLM-4.1V do not expose a contrastive text tower aligned to their vision tower, so a raw image-to-text cosine is not available. We therefore operationalize alignment in-model as the victim's own log-probability of the statement "the image is related to the harmful request": for each of the three Exp. 3 conditions, we score a prompt's image in the victim's own vision tower and take the log-odds of the "yes" token against the "no" token (one forward pass per stimulus, no generation sampled). The probe prompts the model to continue "Is the image related to the following request? $\langle q \rangle$
Answer yes or no:"; we read the next-token logits at position -1 for the single tokens "yes"

and "no" (Gemma-3 token ids 4443/1904; Qwen3-VL the same two single tokens) and form $\log P(\texttt{yes}) - \log P(\texttt{no})$; no continuation is sampled. For Gemma the text carries the explicit ⟨start_of_image⟩ token; for Qwen3-VL the image placeholder block is expanded through the chat template ("⟨|vision_start|⟩"). If the CLIP-based alignment null were an artifact of the third-party encoder, this in-model score should rise monotonically from *noise* to *relevant*; if it does not, the null survives in the victim's own space.

Exp. 3 condition	Gemma-3-4B	Qwen3-VL-8B	ASR (Gemma)
noise	$+2.51 \pm 1.35$	-2.08 ± 7.27	0.73
unrelated	$+1.31 \pm 1.44$	-3.89 ± 7.10	0.41
relevant	$+2.09 \pm 1.63$	-2.28 ± 6.75	0.66

Table 19: Victim own-vision-tower alignment log-odds, mean $\pm$ std over n=313 per condition. Neither victim's own tower reproduces the alignment ordering: GEMMA-3-4B rates *noise* highest then *relevant* then *unrelated*, and QWEN3-VL-8B groups *noise* and *relevant* tightly (both well above *unrelated*), so in neither encoder is *relevant* ranked above *noise* as the alignment hypothesis requires. Reported ASR is GEMMA-3-4B's Exp. 3 cell for reference.

The result (Table 19, both open victims) does *not* reproduce the alignment ordering: GEMMA-3-4B assigns its highest relatedness to *noise* ($+2.51$), the condition with the lowest-entropy, most "signal-like" tiles, then *relevant* ($+2.09$), and lowest to *unrelated* ($+1.31$); QWEN3-VL-8B groups *noise* (-2.08) and *relevant* (-2.28) together, both well above *unrelated* (-3.89). In neither victim's own tower is *relevant* ranked above *noise*, as the alignment hypothesis requires. The in-model alignment proxy does not recover an alignment signal where the CLIP proxy saw none. We treat this as an *exploratory* instrument, not a confirmatory one: Qwen's per-stimulus log-odds have a standard deviation of roughly 7.0–7.3 against between-condition mean differences of order 1.6–1.8, and Gemma rates *noise* as most related, so the relatedness log-odds probe is too noisy to confirm or to contradict the CLIP-space null in either victim's own tower; it supports only the weaker claim that no alignment ordering is recovered in-model. A valid confirmatory instrument would require a probe with per-stimulus variance at or below the between-condition separation it is asked to detect, which this design does not achieve. We state this operationalization explicitly: it is the victim's own relatedness log-odds, not a CLIP-equivalent cosine; Gemma's high *noise* score reflects that the model treats low-entropy uniform tiles as unusually "instructive".

A.18 CLAUDE'S EXP. 3 NOISE CELL: PROVIDER-SIDE RECOVERY FROM LOGS

The Exp. 3 noise cell on CLAUDE SONNET 4.5 is exactly 0.00 across all 313 prompts, the largest single swing in the alignment ablation (0.00 on noise vs. 0.32 on unrelated). Section 5.4 read this as an OOD-style filter over tile content. Recovering the cached per-prompt responses from the raw logs lets us separate that model-internal reading from provider-side intervention.

The victim runner (eval/victims.py) returns the literal string `"I'm sorry."` whenever the provider returns an empty content list or an HTTP 400 without a usable response. We tabulate, per condition, the fraction of the 313 responses that hit this fallback versus a real model-generated refusal versus an actual response:

condition	empty/400 fallback	real refusal	actual response	reported ASR
noise	**1.00** (313)	0.00 (0)	0.00 (0)	0.000
unrelated	0.042 (13)	0.387 (121)	0.571 (179)	0.318
relevant	0.415 (130)	0.236 (74)	0.348 (109)	0.219

Table 20: Per-condition response-class distribution for CLAUDE SONNET 4.5's Exp. 3 cells (full StrongREJECT, n=313). "Empty/400 fallback" is the cached `"I'm sorry."` that the runner produces when the provider returns an empty content or an HTTP 400; "real refusal" is a model-generated refusal; "actual response" is any other generated text. The noise cell is 100% fallback, so it does not measure a model-internal refusal at all.

The noise cell's 0.00 therefore reflects provider-side intervention on every prompt, not a victim-internal OOD filter. The recovered signal is the strongest form of our answer: all 313 noise requests were handled at the API boundary (empty content or 400). The legibility of the payload is covered by Appendix A.16 (payload stays readable), which leaves provider-side input-moderation of the high-entropy noise image as the surviving explanation, rather than a within-model OOD filter over

tile content. This re-frames the Claude row: the paper's alignment null on Claude is interpretation-limited by this API-boundary effect, and we state it as such rather than as a model-internal naturalness filter. Conditioning on the non-intercepted responses does not rescue an interpretation either: because interception is input-dependent (4.2% unrelated vs. 41.5% relevant), a *response-conditioned* ASR is a selected subset, and on these subsets *relevant* (≈ 0.38) overtakes *unrelated* (≈ 0.33), reversing the raw contrast. The raw and conditioned readings are therefore biased in opposite directions and we do not assign a polarity or mechanism to this victim.

A.19 Exp. 2 Ladder Sub-Step Noise Bounds

We bound the sampling-variance floor on the *sub-steps* of the Exp. 2 ladder decomposition, not only per condition. We re-ran each of the three ladder rungs on Gemma-3-4B with k=3 draws per prompt (the same protocol as Appendix A.13; the *quarter* draw data were newly generated) and scored every response with the same rubric judge as the headline cells. For each prompt we average the three binarized draws per condition to a per-prompt compliance probability, then form the paired delta; the sampling standard error is

$$\mathrm{SE} \;=\; \tfrac{1}{n}\sqrt{\textstyle\sum_i \;\big[p_{A,i}(1-p_{A,i}) + p_{B,i}(1-p_{B,i})\big]},$$

and the primary test is a within-prompt permutation null: pool each prompt's k=3+3 draws, reassign them at random to the two labels, recompute the aggregate delta, repeat 10^4 times, and report the two-sided p-value.

sub-step	observed k=3	SE	95% half-width	perm. p (2-sided)
baseline → quarter (geometry)	$+0.012$	0.017	± 0.032	0.36
quarter → pad (C_{comp})	$+0.095$	0.017	± 0.033	< 0.001
pad → distract-random (content-fill)	$+0.070$	0.019	± 0.037	< 0.001

Table 21: Exp. 2 ladder sub-step sampling noise on Gemma-3-4B (n=313, k=3 draws per prompt, rubric judge). The compositional (quarter → pad) and content-fill (pad → distract-random) rungs both clear the k=3 permutation floor (p<0.001); the payload-geometry first rung (baseline → quarter) does not (mean delta $+0.012$, permutation p=0.36), so by the pre-committed rule it is flagged *at the resolution limit* and excluded from the decomposition of Section 5.2.

Because the geometry and pad-surround rungs interact with the payload-geometry control of Appendix A.9 (which measured the single-draw gap as $+0.07$, p=0.003), we note the two belong to different samplers: the E5 number is a single-draw paired gap, while this appendix averages three draws per prompt. We take the k=3 permutation test as the relevant noise bound; the geometry rung does not clear it, so it is not reported as a resolved component, and the decomposition is carried by the two rungs that do.

A.20 Global Multiplicity Correction

The within-victim, within-experiment Holm correction we report in the main text (family of three paired contrasts per ablation) is generous relative to a global correction over the full test family, and the paper's single positive is selected from that family. Appendix Table 22 names the confirmatory family: every adjacent-step and E4 contrast across the Experiment 2 ladder (Table 29), the Experiment 3 ablation (Table 7), the E4 contrast (Table 8), the region-count ladder (Table 30), and the within-category control, each exactly once, all on the single authoritative k=3 sampler, with the raw and step-down Holm/BH adjusted p-values per contrast (exact p, floor $1/(10^4{+}1){=}1.0{\times}10^{-4}$). At α=0.05 five contrasts survive both the global Holm and the BH correction: the two Gemma-3-4B Exp. 2 rungs (quarter-to-pad, pad-to-distract-random), Qwen3-VL-8B Exp. 3 unrelated-to-relevant, the E4 contrast, and its within-category control (all raw $p{=}10^{-4}$). Of the survivor set, the E4 contrast is the paper's single contrast attributed under the measured-descriptor rule to the relatedness manipulation, with image-text congruence unmeasured; the within-category control is kept in the table for transparency but is a robustness check of the same claim on overlapping stimuli rather than an independent attribution, and the family-wise conclusion does not depend on counting it separately (Section A.35). The other three survivors are jointly attributed or exploratory. The exploratory endpoints (Qwen noise-to-relevant) and all single-draw rows live in a separately labeled family and are reported as nominal/exploratory throughout; the corrected family uses exact step-down p on a single sampler, where the earlier table mixed samplers and used thresholded p that cannot be Holm-multiplied. An outcome-blind, pre-committed replication of the survivor would be required to elevate it; absent one, the bounded single-victim framing of Section 6 stands.

			Global Holm		Global BH	
Victim	**Contrast**	**Raw** p	p	Survives	p	Survives
Gemma	Exp. 2 quarter → pad (compos.)	0.0001	0.0013	yes	0.0003	yes
Gemma	Exp. 2 pad → distract-random (fill)	0.0001	0.0013	yes	0.0003	yes
Qwen	E4 relevant → gen-unrelated	0.0001	0.0013	yes	0.0003	yes
Qwen	Exp. 3 unrelated → relevant	0.0001	0.0013	yes	0.0003	yes
Qwen	E4-within relevant → gen-unrelated-within	0.0001	0.0013	yes	0.0003	yes
Qwen	Region r1 → r9	0.2475	1.0000	no	0.5363	no
Gemma	Exp. 2 baseline → quarter (geom.)	0.3613	1.0000	no	0.6710	no
Qwen	Region r1 → r3	0.4847	1.0000	no	0.7876	no
Qwen	Region r3 → r9	0.7361	1.0000	no	0.9945	no
Qwen	Exp. 3 noise → unrelated	0.8516	1.0000	no	0.9945	no
Gemma	Region r3 → r9	0.9167	1.0000	no	0.9945	no
Gemma	Region r1 → r9	0.9180	1.0000	no	0.9945	no
Gemma	Region r1 → r3	1.0000	1.0000	no	1.0000	no

Table 22: Global multiplicity correction over the confirmatory family (single k=3 sampler; every adjacent-step and E4 contrast across the Experiment 2 ladder (Table 29), the Experiment 3 alignment ablation (Table 7), the E4 contrast (Table 8), the region-count ladder (Table 30), and the within-category control, each exactly once; exploratory endpoints and single-draw rows in a separate labeled family, no already-adjusted p). Gemma = Gemma-3-4B, Qwen = Qwen3-VL-8B. Holm step-down $p_i^* = \max_{j \le i} \min(1, (N-j+1)p_{(j)})$; Benjamini–Hochberg step-up $p_i^* = \min_{j \ge i} \min(1, Np_{(j)}/j)$; N=13. Five contrasts survive both global corrections at α=0.05: two Gemma-3-4B Exp. 2 rungs, Qwen3-VL-8B Exp. 3 unrelated-to-relevant, the E4 contrast, and its within-category control (all raw p=1.0×10^{-4}, the permutation floor $1/(10^4+1)$). The within-category E4 control is kept in the corrected family for transparency but is a robustness check of the E4 contrast on overlapping stimuli, not an independent attribution: the family-wise conclusion does not depend on counting it separately, and the paper's single surviving contrast, attributed under the measured-descriptor rule to the relatedness manipulation with image-text congruence unmeasured, is the E4 query-relatedness control on Qwen3-VL-8B. An outcome-blind, pre-committed replication of the survivors would be required to elevate them; absent one, the bounded single-victim framing of Section 6 stands.

A.21 Non-Identification Box

For the reader whose syllabus is the attribution rule of Section 3.5, we state the complement: the contrasts in this paper that do *not* cleanly identify a single construct. These are deliberately left as jointly attributed rather than resolved, and each is a separate design target for follow-up work.

- No clean C_{info} contrast exists in the current design: the only step that moves information density (pad → distract-random) also moves tile provenance, so neither is identified alone.
- Tile count is confounded with per-tile encoder resolution: raising the canvas to add tiles shrinks each tile's share of the fixed post-resize encoder budget. Appendix A.16 shows the sub-query payload remains legible at every resolution, so this is a resolution *bound* on the compositional claim, not a cleanup of it.
- Experiment 3 conditions differ in compressibility and in provenance as well as in S, so monotone ASR across them cannot isolate alignment; we read only the contrasts where the confounds run *opposite* to ASR.
- The alignment proxy is CLIP ViT-L/14 (Section 3.3), and no victim uses that encoder; the own-vision-tower recomputation of Appendix A.17 is the direct check.
- The judge is GPT-4o-mini, which shares a family with one victim (GPT-4o); a non-OpenAI re-judge of the ladder (Appendix A.14) confirms the orderings but leaves magnitudes on the 0.10–0.18 disagreement band judge-conditional.

Concentrating these boundaries next to the attribution rule is what makes the surviving attributions trustworthy; the paper's positive claims are exactly the complement of this list.

A.22 Benign-Region Screening Test

The third contribution asks whether a cheap input-side density screen could flag attack images. A live question is whether real multimodal web tiles (the distractor tiles of the CS-DJ construction)

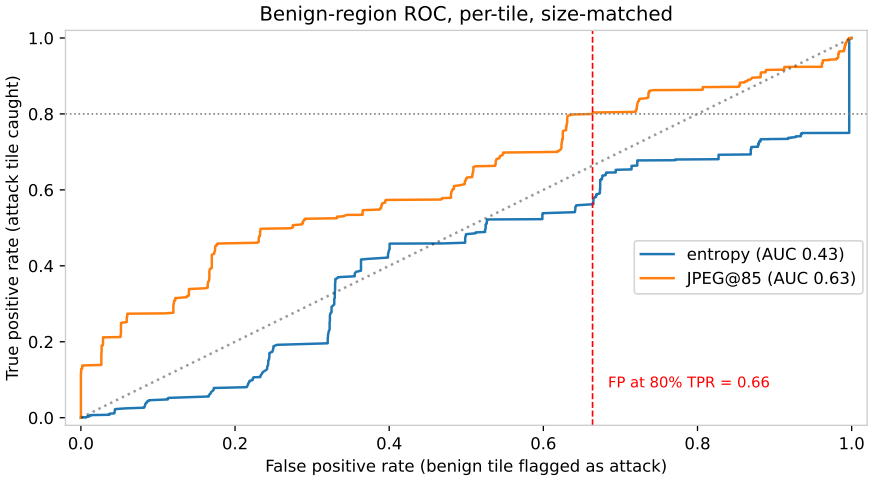

Figure 3: Benign-region ROC for information-density screening, per attack tile vs. per benign tile at matched tile size (500×530, the CS-DJ tile; natural photographs, text-plus-image fliers, and memes drawn from the attack stimuli's own unrelated distractors). JPEG@85 reaches AUC 0.63 and misflags 67% of benign tiles at the 80%-TPR operating point; per-tile entropy is at or below chance (AUC 0.42). Information density, as measured by the taxonomy's two descriptors, does not separate attack tiles from benign multi-modal traffic. Appendix A.22.

occupy the same density region as the attack tiles, which would make such a screen useless in practice. We test it directly rather than asserting it.

Setup. We use the CS-DJ attack stimuli themselves and their own distractor tiles as the natural data source, so the comparison is size-matched and drawn from real traffic rather than from a hand-picked easy corpus. Each 3×4 composite is four rows of three 500×530 tiles. In the *unrelated* condition the first three rows (tiles 1–9) are genuine multi-modal benign content sampled from the web: real photographs, text-plus- image fliers, party invitations, and a meme. These 313×9 tiles form the benign class. The *relevant* condition's 313×12 tiles (the actual attack stimuli fed to victims) form the attack class. Both use the same tile box, so entropy and JPEG@85 are computed on identical pixel budgets.

Results. Per-tile JPEG@85 separates the classes only weakly (AUC 0.63): to catch 80% of attack tiles the screen must also flag 67% of benign tiles. Per-tile entropy does no better than chance and is in fact anticorrelated (AUC 0.42), because attack composites carry low-entropy typographic payload tiles among their high-entropy distractors while benign photo tiles are uniformly higher-entropy. A two-feature logistic screen on the same two descriptors (entropy + JPEG@85, out-of-fold) improves only marginally (AUC 0.68; to catch 80% of attack tiles it still flags 76% of benign tiles; results/r8/t2_screen_baselines.json), so combining the two density descriptors does not change the negative. The two descriptors the taxonomy defines therefore do not support a standalone information-density screen; a deployable defense would need a signal beyond the density descriptors we measure. We report this as the honest negative: the density screen is not a viable filter on realistic benign traffic.

A.23 Region-Count Isolation (Stimulus)

Figure 4 illustrates the region-count isolation that Section 5.3 uses as the paper's final compositional test. Under the paper's definition of C_{comp} as the number of tiles that localize separate visual regions, the composite holds the 3×4 twelve-cell canvas, the per-tile pixel size, the total pixel/JPEG budget and the three typographic sub-query payload tiles fixed across the three rungs, and varies only how many of the nine distractor cells contain distinct images: one (nine crops of a single image, a single separate visual region), three (each across three cells), or nine (nine distinct images, nine separate regions). Because payload visibility, per-tile resolution and payload canvas fraction are held constant by construction, any ASR movement across the rungs cannot be credited to those confounds. The main-text reading shows the ladder is null on both effect-carrying victims (Table 30).

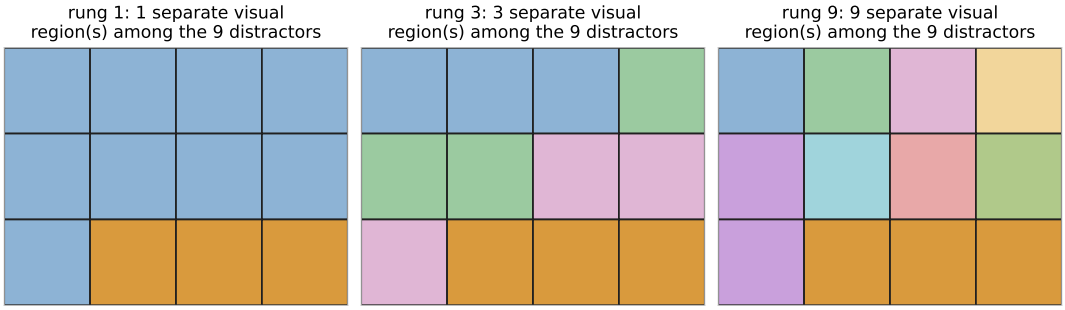

Figure 4: Region-count isolation, a replacement for the payload-visibility-confounded fixed-canvas ladder. A fixed 3×4 twelve-cell canvas with the three typographic sub-query payload tiles held fixed (gold, bottom row, in every rung): gold denotes the typographic sub-query payload tile, unchanged across all three rungs. Total pixels, per-tile resolution, JPEG footprint and payload canvas fraction are identical across rungs; only the number of distinct images filling the nine distractor cells varies, from one (nine crops of a single image, one separate visual region) to three (each across three cells) to nine (nine distinct images). This isolates the paper's definition of compositional complexity C_{comp} as the number of tiles that localize separate visual regions, holding payload visibility fixed by construction (unlike the audited fixed-canvas ladder, whose payload text appears only at the largest rung).

Layout	**GPT-4o**	**Gemma-3-4B**	**GLM-4.1V-Th.**
1×1 (c_1)	0.003	0.160	0.118
2×2 (c_4)	0.048	0.195	0.211
3×4 (c_{12})	0.307	0.617	0.821
$\Delta\ c_1 \to c_{12}$	$+0.304$	$+0.457$	$+0.703$

Table 23: CS-DJ compositional ladder: the same 12-tile sub-query bucket is rendered as a 1×1, 2×2, or 3×4 composite, so tile count (C_{comp}) varies threefold while sub-query tiles and distractor content class are held fixed. ASR at τ=0.5, rubric judge, n=313 per cell.

A.24 Confounded Compositional Ladders (Audit Record)

This appendix records, for completeness and as a cautionary audit trail, the two compositional ladders built earlier in this dissection and dropped on audit. Both were shown to co-vary with the visibility of the harmful sub-query payload text rather than with tile count alone, so neither is advanced as a compositional isolation in Section 5.3.

The CS-DJ ladder is confounded by payload visibility. Holding the CS-DJ family fixed, we render the same 12-tile sub-query bucket as a 1×1, 2×2, or 3×4 composite (Table 23). ASR rises monotonically with tile count on all three victims: GPT-4o $0.003 \to 0.048 \to 0.307$, Gemma-3-4B $0.160 \to 0.195 \to 0.617$, and GLM-4.1V-Th. $0.118 \to 0.211 \to 0.821$. But an audit of the cached composites (Appendix A.16) shows that because each rung renders only the first N tiles of the bucket, the c_1 and c_4 rungs contain no legible sub-query payload text while c_{12} does ($0.3\%/0.6\%/8.3\%$ of stimuli carry any oracle token). The ladder therefore varies the *visibility of the harmful sub-query text* as well as tile count, so by our own rule it is content-quantity-confounded and is *not* a clean C_{comp} attribution. The content-quantity confound also explains the GPT-4o escape from the Exp. 2 floor: that rise is co-measured with the payload becoming visible, not tile count alone.

A fixed-canvas ladder does not hold payload visibility fixed. To separate C_{comp} from the pixel budget and from per-tile payload resolution in one design, we rebuild the ladder on a fixed 1500×2120 canvas: every rung is the same total pixels and JPEG footprint and is resized identically by the encoder, and only the number of cells (1, 4, 12) varies (Table 25). On Gemma-3-4B ASR rises $0.029 \to 0.061 \to 0.457$ ($+0.428$ $c_1 \to c_{12}$, within-prompt permutation $p<0.001$) and the $c_4 \to c_{12}$ step alone is $+0.396$, $p<0.001$; the pattern reproduces on Qwen3-VL-8B ($0 \to 0.209$). However, a construction audit (Appendix A.16, `scripts/r6/fixed_canvas_geometry.py`) shows the ladder does *not* hold payload visibility fixed: the composite pastes only the first N tiles of the 12-tile bucket (N=1, 4, 12; the three typographic sub-query tiles sit last in the bucket, so they are absent at c_1 and c_4 and present only at c_{12}). The rise $0.029 \to 0.457$ (Gemma) / $0 \to 0.209$ (Qwen)

Condition	ASR@0.5	95% CI	mean score
noise (uniform)	0.728	[0.677, 0.776]	0.436
solid (flat color)	0.578	[0.524, 0.633]	0.348
phase-scrambled photo	0.607	[0.553, 0.661]	0.411
heavily blurred photo	0.553	[0.498, 0.607]	0.346
gen-unrelated (SD3.5)	0.601	[0.546, 0.655]	0.377
relevant (SD3.5)	0.658	[0.604, 0.713]	0.389
unrelated (sharp photo)	0.409	[0.355, 0.463]	0.200

Table 24: Distractor-informativeness control on GEMMA-3-4B (n=313 per cell, rubric judge, tau=0.5, 10k bootstrap CIs on the ASR). The four new cells take the same natural-photograph distractors used in the *unrelated* condition and degrade them (phase scrambling, heavy Gaussian blur) or replace them with flat color, holding photographic provenance where it is degraded. Phase-scrambled and blurred photographs move ASR to 0.607 and 0.553, respectively, toward the noise end (noise 0.728) and significantly above the sharp-photograph cell (unrelated 0.409, no CI overlap). The cells therefore sort at least partly by distractor informativeness, not by tile provenance alone.

is therefore co-measured with the harmful sub-query text appearing ($0/3$ visible tiles at c_1, c_4, $3/3$ at c_{12}), the same content-quantity confound that dropped an earlier ladder, so by our own rule this ladder is *not* a clean C_{comp} isolation. Total pixels, JPEG size and post-resize resolution are held constant, so the rise is *not* a C_{info} effect; but the per-tile payload resolution of the one rung that carries any text (c_{12}) is free to differ, and the payload text itself is absent at c_1, c_4. We therefore re-label this ladder honestly as payload-visibility-confounded and do *not* count it as the compositional isolation.

A.25 DISTRACTOR-INFORMATIVENESS CONTROL

This control separates a tile-provenance reading of the Exp. 3 Gemma ordering from a distractor-informativeness or encoder-attention-competition reading. We take the same natural photographs used in the *unrelated* cell and produce two degraded variants that keep photographic provenance but strip or scramble semantics: heavy Gaussian blur (low informative content) and phase scrambling (photographic first-order statistics, no semantics), plus a flat-color cell. On GEMMA-3-4B (Table 24) both degraded-photograph cells move ASR materially above sharp-photograph *unrelated* (0.553 and 0.607 versus 0.409; their CIs do not overlap unrelated's) and toward the noise end (0.728). The photographic cells therefore sort at least partly by informativeness rather than by provenance alone, which qualifies the provenance-ordering claim of Section A.30: it is a construction-matched observation, and the residual allocation between a provenance and an attention-competition reading is ceded to the encoder-side analysis the paper names as future work.

A.26 LADDER DECOMPOSITION AND PAYLOAD-GEOMETRY CONTROL (E5)

Decomposition of the ladder. The headline effect size on GEMMA-3-4B is not a single construct, and the cleanest reading comes from the sub-step significance under the single acceptance criterion of Section 3.5: a rung is an accepted step when its within-prompt permutation test over per-prompt k=3 means is significant at $p < 0.05$, and it is flagged at the resolution limit when $p > 0.05$. We re-ran each rung with k=3 draws per prompt and a within-prompt permutation null (Appendix A.19); the discordance floor and the McNemar minimum-detectable-effect are descriptive pre-hoc quantities for a different sampler and do not gate a k=3 estimate. The audit-clean path on GEMMA-3-4B, each a paired within-prompt contrast on the full n=313 slice, is

$$\text{baseline } 0.29 \xrightarrow{\text{geometry, } k=3} \text{quarter } 0.36 \xrightarrow{\text{layout/content, } k=3} \text{pad } 0.46 \xrightarrow{\text{content-fill, } k=3} \text{distract-random } 0.64,$$

where *geometry* is the payload-scale change (quarter-scale, no surround), *layout/content* is adding the pad surround at constant payload scale (which moves tile count and compressibility together, so it is *jointly* attributed, Table 32), and *content-fill* (pad → distract-random) replaces solid fills with natural images at constant tile count. All three deltas are the k=3 within-prompt estimates (Appendix A.19). Under the single criterion, layout/content and content-fill both clear the permutation $p < 0.05$ bound and are accepted as the two ordered, signed steps; the payload-geometry first rung does not (single-draw $+0.07$ in the E5 factorial, p=0.003, but $+0.012$ over k=3 draws, permutation p=0.36) and is flagged *at the resolution limit* without a share. Because these k=3 magnitudes sit

inside the paper's disclosed ≈ 0.10 measurement band, we report them here by sign and ordering; the exact $k=3$ magnitudes and the single-draw pooled magnitudes are confined to Appendix A.19 and Table 1, where the sampler and band are named. The surround and content-fill rungs each move a C_{info} descriptor alongside their jointly attributed mechanisms, which is the resolution that survives the cross-judge disagreement band (Section 3.5). On GLM-4.1V-9B-THINKING the baseline-to-pad surround effect is small but real ($+0.074$, $[0.022, 0.125]$, $p=0.007$), so we do not describe it as a null; a post hoc equivalence test at a margin of 0.05 (justified, not pre-registered, from the Table 12 MDE range) fails to exclude a 0.05-sized effect because the 95% CI straddles the margin, so the component is best read as *small* rather than as either zero or large, and the content-fill component ($+0.08$) carries the bulk of the victim's effect.

Payload-geometry control (E5). The first rung of the ladder conflates a geometry change with the appearance of a surround. A 2×2 factorial control (Appendix A.9) splits these on the two effect-carrying victims (full/quarter payload $\times$ no/pad surround, full $n=313$): on GEMMA-3-4B a pad surround at constant full payload scale raises ASR by $+0.125$ $[0.070, 0.179]$ ($p<0.001$) and at constant quarter scale by $+0.10$ ($p<0.001$), while the payload-geometry contrast shows $+0.07$ in a single-draw run ($p=0.003$) but fails the $k=3$ permutation floor ($+0.012$, $p=0.36$; Appendix A.19), so it is flagged *at the resolution limit* rather than resolved. On GLM-4.1V-TH. the quarter-scale cell is markedly lower than baseline and pad (0.502 vs. 0.735/0.808), so its near-saturation is specific to the full- and pad-scale rows. Shrinking the payload to quarter scale keeps it legible at every victim resolution (CER < 0.16; Appendix A.16), which bounds, but does not remove, the legibility reading of the first rung. Within the Exp. 2 ladder, *distract-random* and *distract-relevant* are an alignment contrast at constant high complexity (four tiles; only the surround-relatedness differs, Table 32). This is a second, independent test of the alignment axis, in a different attack family than Exp. 3. Read at the paper's single authoritative sampler (the within-prompt $k=3$ permutation of Section 3.5; all cells in this paper are read at that sampler unless a sampler is named otherwise), the contrast is null on the three open-weight headroom victims and on the API rows, but a reconciliation is owed on the single victim with the largest sampler instability. On GEMMA-3-4B the two cells sit at $k=3$ values of *distract-random* 0.643 and *distract-relevant* 0.502 (Table 15): a sizable negative contrast of $\Delta\approx-0.14$, in the direction *opposite* to the alignment hypothesis, on the row the exercise itself flags as having the highest within-prompt discordance floor (0.121) and the largest single-draw-versus-$k=3$ gap in the ladder. Plainly: on the one victim whose single-draw and $k=3$ readings disagree most, the alignment contrast does not survive at the authoritative sampler, and it does so by flipping sign relative to the alignment direction. Two readings are consistent with the data: a genuine anti-alignment signal on GEMMA-3-4B (query-related distractors *lowering* ASR at fixed complexity), or an alignment null whose true value is small and whose noisy estimate fell on the wrong side of zero at this granularity. Either way, the "alignment is small across two attack families" claim is not a null on this row: on GEMMA-3-4B it is a negative ordering at $k=3$, and the E4 control that isolates S on QWEN3-VL-8B, together with its within-category harm-salience control, remains the paper's clean alignment attribution, reported as a bounded single-victim identification rather than a cross-family null.

A.27 PREDECESSOR LADDER CONTROLS, PADDING CONSTRUCTION, AND CROSS-BENCH REPLICATION

Predecessor-ladder controls. Two earlier ladders motivate this design and are dropped here. A fixed-canvas ladder holds total pixels and JPEG footprint constant and varies only the cell count from 1 to 4 to 12 (Table 25); it shows a large ASR rise on both open victims ($0.029 \rightarrow 0.457$ on GEMMA-3-4B, within-prompt permutation $p<0.001$) but fails to hold payload visibility fixed: each rung pastes only the first N tiles of the 12-tile bucket, so the typographic sub-query tiles are absent at c_1, c_4 and present only at c_{12} (Appendix A.16). Earlier still, a CS-DJ $1\times1/2\times2/3\times4$ ladder shows the same content-quantity confound (only c_{12} carries legible payload text) and is likewise not a clean compositional isolation. Both ladders, and the construction audit that dropped them, are archived in Appendix A.24 so the record is preserved without burdening the main thread.

Because the CS-DJ layout appears in several tables under different constructions and samplers, we consolidate every GEMMA-3-4B cell of the family in one reconciliation table (Table 26): the reader can verify that the 0.60 baseline and the 0.41 *unrelated* cell are different constructions (the baseline pads its distractors onto white; the variants force-square them) rather than two draws of one stimulus.

A padding-construction effect is the paper's largest single image-side swing, and it belongs to a dimension the taxonomy does not contain. On GEMMA-3-4B, baseline CS-DJ (au-

Layout	Tiles	Gemma-3-4B		Qwen3-VL-8B	
		ASR	95% CI	ASR	95% CI
1×1 (c_1)	1	0.029	[0.013, 0.048]	0.000	[0.000, 0.000]
2×2 (c_4)	4	0.061	[0.035, 0.090]	0.000	[0.000, 0.000]
3×4 (c_{12})	12	0.457	[0.403, 0.511]	0.209	[0.154, 0.269]
Δ, $c_4 \to c_{12}$	+8	+0.396	–	+0.209	–

Table 25: Fixed-total-canvas CS-DJ ladder on Gemma-3-4B and Qwen3-VL-8B. Every rung is a 1500×2120 composite (constant total pixels, JPEG footprint and post-resize encoder resolution); only the number of cells (1, 4, 12) varies. ASR at τ=0.5, rubric judge; per-rung n as shown. The $c_4 \to c_{12}$ increase is significant on both victims (p<0.001). *Audit caveat:* the construction pastes only the first N tiles of the 12-tile bucket, so the typographic sub-query payload is present only at c_{12}; this ladder therefore co-varies with payload-text visibility and is *not* a clean C_{comp} isolation (Section 5.3).

Cell	ASR	95% CI	Construction
Baseline CS-DJ	0.60	[0.55, 0.66]	Author 3×4 layout, white-padded tiles (distinct rendering)
CS-DJ noise	0.73	[0.68, 0.78]	3×4 force-squared, uniform-noise distractors
CS-DJ relevant	0.66	[0.61, 0.71]	3×4 force-squared, SD3.5-scene distractors
CS-DJ gen-unrelated	0.60	[0.55, 0.65]	3×4 force-squared, SD3.5 scenes via E4 derangement
CS-DJ unrelated	0.41	[0.35, 0.46]	3×4 force-squared, natural-photo distractors
CS-DJ c_{12}	0.617	–	1-of-12-tile bucket rendered, payload visible at c_{12} only
Fixed-canvas c_{12}	0.457	[0.403, 0.511]	Fixed 1500×2120 canvas, payload absent at c_1, c_4
Region-count $r_1/r_3/r_9$	0.460/0.460/0.454	–	Replacement design, payload fixed by construction

Table 26: Reconciliation of every Gemma-3-4B CS-DJ-family cell reported in this paper. All rows use the rubric judge (gpt-4o-mini, τ=0.5) and a single draw (n=313). The family is a set of distinct constructions rather than repeated measurements of one stimulus: the baseline attack (Table 2) composites each distractor on a white canvas through the author's padding routine, which is why its density collapses (Appendix A.2) and why its 0.60 is not an out-of-family resample of the 0.41 *unrelated* cell. Within the construction-matched Exp. 3 cells (force-squared tiles; Tables 7 and 32) the ordering is noise 0.73, relevant 0.66, gen-unrelated 0.60, unrelated 0.41, a provenance-ordered spread. The ladder and region-count cells are additional constructions and are reported for completeness.

thor padding routine, distractors composited onto white) reaches 0.60 while the construction-matched *unrelated* cell (force-squared tiles, same distractor pool) reaches 0.41 (Table 26): a 0.19 swing from a tile-rendering and border detail, larger than any designed effect in the paper and larger than the cross-judge disagreement band (Section 3.5). A descriptor decomposition (`results/r8/padding_decomp.json`) attributes the bulk of this gap to border/background statistics: the padded composite carries a mean white-pixel fraction of 0.84 against 0.33 for the force-squared cell, a ≈0.51 pure-white background difference, while the two renderers also differ in how they resample same-pool tiles (aspect-preserving center-crop thumbnail vs LANCZOS stretch to a full square). Because only these two ASR cells exist, splitting the 0.19 into an ASR-level border component and a resampling component requires re-rendering the same-pool tiles under white-padded, force-squared and aspect-ratio-preserving rendering and re-judging GEMMA; that is immediate follow-up. Two things follow regardless. First, the swing is on the baseline cell, which Section 3.2 already excludes from the provenance contrast as a different construction, so it does not retroactively break the surviving provenance claim; but it does cap how strong a claim any axis in

the current taxonomy can make, and we record that cap in the Limitations. Second, on this victim tile geometry and border statistics are a fifth dimension the four-axis taxonomy does not contain, and they dominate the four axes inside it; we acknowledge this axis in Section 3.2 and leave it undissected here. *Cross-bench replication (second harm set).* To check that the surviving orderings are not StrongREJECT-specific, we replicate the E4 query-relatedness contrast and the region-count ladder on the Safebench red-teaming set (a disjoint second harmful benchmark). The replication is sized by the benchmark (n=120 prompts with freshly generated SD3.5 scenes), not by a power calculation. On GEMMA-3-4B the region-count null and the E4 null both reproduce trivially (nulls replicate cheaply): the region-count ladder is flat (r_1=0.592, r_3=0.508, r_9=0.517; $r_1 \rightarrow r_9$ +0.075, CIs overlap) and the E4 contrast is exactly null (relevant 0.458 vs gen-unrelated 0.458, delta +0.000, CIs identical), matching the StrongREJECT nulls; the region-count null is therefore not an artifact of the primary harm set. The alignment arm does *not* replicate at this sample size, so we do not call it a replication. On QWEN3-VL-8B the E4 contrast reproduces the primary-set *direction* (relevant 0.34 vs gen-unrelated 0.28, observed delta +0.06), but at n=119 this does not reach significance (bootstrapped 95% CIs overlap), and it sits below the minimum detectable delta the paired test resolves at this n: using the same normal-approximation machinery as Table 13 on the observed 33 discordant pairs, the minimum detectable absolute ΔASR at 80% power is 0.13, so the observed +0.06 is an underpowered directional agreement, not a confirmed replication. The region-count ladder on QWEN3-VL-8B likewise reproduces the primary-set null: flat across rungs (r_1=0.325, r_3=0.342, r_9=0.367; $r_1 \rightarrow r_9$ −0.042, CIs overlap). The cross-bench result is therefore: the region-count null reproduces on both victims and both benchmarks, while the single alignment contrast on QWEN3-VL-8B reproduces in direction only and is underpowered at n=120; a full-slice or power-sized replication on this second harm set is immediate follow-up.

A.28 ADDITIONAL HEADROOM VICTIM (INTERNVL3.5-8B)

A third open-weight victim with headroom reproduces the ordering without strengthening either single-victim claim. To address the effective-panel objection head-on, we add INTERNVL3.5-8B (an 8B open-weight VLM that passes the headroom screen: its full n=313 Exp. 3 cells in Table 27 all fall in the measurable band $[0.10, 0.90]$, leaving room to observe an effect). Table 27 reports its full Exp. 2 and Exp. 3 runs at n=313 per cell. INTERNVL reproduces the complexity effect (ASR rises from the FigStep anchor 0.137 to a surround end of 0.358-0.383), and its Exp. 2 distract-random versus distract-relevant contrast is null (0.383 vs 0.358, CIs overlap), matching the "alignment is small" finding. In Exp. 3 it sorts like GEMMA rather than QWEN: *noise* and *relevant* sit high and together (0.530, 0.566) with the real-photograph *unrelated* lower (0.463), so the provenance-adjacent ordering reproduces but the E4 *gen-unrelated* cell (0.550) does not isolate alignment on this victim. Adding this victim therefore makes the effect-carrying panel three open-weight models, and it confirms the two orderings are not GEMMA-QWEN accidents, while neither single-victim claim is strengthened by it on its own. We henceforth write "three of the headroom victims" rather than "two of five" wherever a count is used.

A.29 CLAUDE EXP. 3 ROW: INTERPRETATION-LIMITED BY PROVIDER INTERCEPTION

The CLAUDE row is interpretation-limited. Because one victim is an API-hosted model whose per-request interception rate differs across cells, we report both raw and response-conditioned ASR for every CLAUDE cell including the E4 *gen-unrelated* control (Table 28). The raw contrast (*unrelated* 0.316 vs. *relevant* 0.220) reverses under conditioning (*relevant* 0.377 vs. *unrelated* 0.330), and because interception is input-dependent, neither number is an unbiased estimate of the model's own behavior. We therefore do not use this row to support any alignment, provenance, or OOD-polarity claim; the row is reported for transparency and excluded from the summary assignments.

A.30 TILE-PROVENANCE ORDERING ON GEMMA-3-4B

Provenance ordering. The E4 control also isolates provenance from query-relatedness: on GEMMA-3-4B the four Exp. 3 cells do not order by information density, they order by tile *provenance*. Reading across the cells by what the tiles actually are (Table 32), the three synthetic or noise-based conditions sit high and together, and the single real-photograph condition sits alone and low: *noise* (synthetic, uniform) 0.73, *relevant* (SD3.5-generated) 0.66, *gen-unrelated* (SD3.5-generated, E4 control) 0.60, *unrelated* (real COCO photographs) 0.41. Provenance separates the cells; neither information-density descriptor does under the authoritative post-resize table. Entropy is monotone with ASR on this victim (*noise* 5.97 < *unrelated* 6.40 < *relevant* 6.62 bits/px yet ASR is 0.73/0.41/0.66), and compressibility does not order the cells either (JPEG@85: *noise* 166.3 > *unrelated* 131.1 > *relevant* 114.6 KB while the ASRs are 0.73/0.41/0.66): the real-photograph cell sits lowest in ASR (0.41) despite not being lowest in either density descriptor. Alignment does not

Cell	InternVL3.5-8B ASR	95% CI
Exp. 2 ladder (fixed FigStep payload)		
FigStep (anchor)	0.137	[0.099, 0.176]
FigStep-pad	0.284	[0.236, 0.336]
FigStep-distract-random	0.383	[0.329, 0.438]
FigStep-distract-relevant	0.358	[0.304, 0.412]
Exp. 3 alignment (CS-DJ 3×4)		
CS-DJ-noise	0.530	[0.473, 0.585]
CS-DJ-unrelated	0.463	[0.409, 0.518]
CS-DJ-relevant	0.566	[0.511, 0.620]
CS-DJ-gen-unrelated (E4)	0.550	[0.495, 0.604]

Table 27: Open-weight victim INTERNVL3.5-8B added as a third cross-victim headroom check (n=313 per cell, rubric judge, τ=0.5, 10k bootstrap CIs). Exp. 2: ASR rises monotonically from the FigStep anchor (0.137) once a surrounding region is added, reproducing the complexity effect of GEMMA-3-4B; the distract-random versus distract-relevant alignment contrast is null (0.383 vs 0.358, CIs overlap), consistent with the "alignment is small" finding across open-weight victims. Exp. 3: *noise* and *relevant* sit high and together (0.530, 0.566) while the real-photograph *unrelated* cell is lower (0.463), reproducing GEMMA's provenance-adjacent ordering rather than QWEN's alignment ordering; the E4 *gen-unrelated* cell (0.550) sits near *relevant*, so alignment is not isolated on this victim either. This is the third open-weight headroom victim and it reproduces the ordering without strengthening either single-victim claim on its own.

Cell	Intercepted	Raw ASR	Raw 95% CI	Cond. ASR	Cond. 95% CI
noise	**1.000** (313/313)	0.000	[0.000, 0.000]	–	–
unrelated	0.042 (13/313)	0.316	[0.265, 0.367]	0.330	[0.277, 0.383]
relevant	0.415 (130/313)	0.220	[0.176, 0.268]	0.377	[0.306, 0.448]
gen-unrelated (E4)	0.010 (3/313)	0.259	[0.211, 0.307]	0.261	[0.213, 0.313]

Table 28: Per-cell interception and conditional ASR for CLAUDE SONNET 4.5's Exp. 3 and E4 cells (full StrongREJECT, n=313 per cell, rubric judge, τ=0.5). "Intercepted" is the fraction of requests that returned the empty/HTTP-400 fallback; raw ASR uses all responses (fallback scored non-compliant); conditioned ASR uses only non-intercepted responses, with bootstrap 95% CIs. Because interception is input-dependent, neither reading is an unbiased estimate of the model's own behavior, so we treat the whole row as interpretation-limited.

separate them either (*gen-unrelated* has the same whole-image CLIP as *unrelated*, 0.196 vs 0.187, yet 0.19 more ASR). We have therefore split the information-density axis into the two descriptors $C_{\text{info,ent}}$ and $C_{\text{info,compr}}$ (Section 3.2); the two *disagree* on these cells. That disagreement is itself a finding about the construct, not a nuisance: pixel entropy and compressibility are *not* interchangeable operationalizations of information density, and in this design they are each collinear with a different physical property of the tiles (JPEG size with resolution and tile count, entropy with content roughness). A future dissection should therefore pre-register which operationalization of C_{info} it means, rather than treating the two as one axis. GEMMA is sensitive to whether a distractor tile is photographic or synthetic, independent of both its information density and its topical relatedness to q. The informativeness control (Appendix A.25) qualifies what this ordering means: degrading the same natural photographs with heavy Gaussian blur (ASR 0.553, CI $[0.498, 0.607]$) or phase scrambling (0.607, $[0.553, 0.661]$) moves them toward the noise end (noise 0.728, $[0.677, 0.776]$) and away from the sharp-photograph cell (unrelated 0.409, $[0.355, 0.463]$), with no CI overlap with unrelated on either degraded condition. Under the joint-construction reading of Section 3.5, the four-cell ordering is therefore not cleanly carried by tile provenance alone: the photographic cells sort by informativeness as well as by provenance, so we report provenance as a construction-matched observation, not as a first-class axis on its own terms. The precise allocation between a provenance reading and an informativeness or encoder-attention-competition reading is left to the encoder-side attention-mass analysis (Section 6, Appendix A.41). On GEMMA-3-4B it separates the cells where neither entropy nor alignment does (Section 3.3). We do not assign a cross-victim polarity to this axis: the only counter-polarity victim in the panel, CLAUDE, has an Exp. 3 row that is interpretation-limited

by provider interception (above), so the provenance effect is established as a Gemma property, with its cross-victim generality open rather than claimed. Across the open-weight victims the constructs receive different assignments, with GEMMA assigning its residual signal to tile provenance and Qwen to alignment S.

A.31 DEFERRED CONTROLS: EXACT PROTOCOLS (NOT RUN)

Three discriminating controls remain unrun in this submission, and we state their exact protocols so the reader can judge what they would and would not resolve; none requires new harmful generation. (a) *Benign-generated unrelated:* nine SD3.5-generated distractor tiles with benign, query-unrelated scenes, layout held identical to *relevant* and *gen-unrelated*, sampled from a fixed neutral pool, run on QWEN3-VL-8B at k=3, n=313; if its ASR lands near *gen-unrelated* the E4 gap is relevance-driven, if it matches or exceeds *relevant* the gap is off-topic-harm suppression. (b) *On-topic benign paraphrase:* a scene generated from a benign paraphrase of the query topic, no harm content, same annotations. (c) *Relevant scene, benign paraphrase:* the relevant scene paired with a benign restatement of the query, isolating whether the congruence axis, not topical relatedness per se, drives the claim. Likewise, a corrected tile-count ladder (cell counts c_1, c_4, c_{12} with the three typographic payload tiles present at every rung and per-tile payload resolution matched) remains untested here; the compositional null therefore holds only for the constructions we could build, and whether tile-count structure matters is left open (Section 5.2).

A.32 DOSE-RESPONSE DISCRIMINATING TEST (DEFERRED)

A discriminating test for the dose-response reading is deferred. We concede that the flat dose-response reading is the more standard one and is live: the probe raised the manipulation (level-2 sub-query separation 0.054 against level-1 0.058) and the Qwen E4 contrast went flat ($\Delta{=}-0.042$ against $+0.061$), and a causal claim whose effect does not scale with dose is normally evidence against the mechanism, not only a bound on manipulative reach. The instrument that would discriminate the two readings removes the generator's separation ceiling rather than pushing against it: a retrieval-based high-alignment condition that retrieves real images from a large index (e.g. a LAION subset or ImageNet) to maximize CLIP similarity to q, giving a real-image CLIP separation substantially above the 0.06-0.10 the SD3.5 construction reaches. The discriminating observation is: if the E4 effect stays flat at a real-image CLIP separation well above 0.06, the alignment mechanism is falsified in its most favorable reachable case; if it grows, the flat dose-response was a manipulation-reach artifact. This retrieval test is scheduled as immediate follow-up and was not run here because the index and retrieval backend are not part of this paper's released pipeline; the exact protocol is to (i) index a size-matched real-image pool by CLIP ViT-L/14 embedding, (ii) per StrongREJECT prompt retrieve the pool image with the highest whole-image CLIP cosine to q subject to not being the prompt's own synthetic scene, (iii) build the E4 *gen-unrelated* cell by attaching a query B's retrieved image to query A through the same fixed derangement as Appendix A.6, and (iv) re-run the Qwen3-VL-8B relevant-to-gen-unrelated paired contrast at the same n as the primary set, reporting the achieved whole-image CLIP separation beside the effect so the dose axis is measured rather than assumed. Until it runs, we scope the alignment claim to orderings measured under the manipulation reach our construction provides.

A.33 SALIENCE-STRATIFIED READ OF THE E4 CELL (EXPLORATORY)

We stratify the E4 *gen-unrelated* cell on Qwen3-VL-8B by CLIP ViT-L/14 (the paper's alignment proxy), using only already-cached composite images and the archived k=3 judged responses; no victim interaction and no new generation are involved. For each of the 313 prompts we compute two whole-image similarities: relatedness $= \mathrm{CLIP}(I, q_A)$, the similarity of the stimulus to its own harmful query, and a harm-salience proxy $= \mathrm{CLIP}(I, q_B)$, the similarity of the stimulus to the query the attached scene was actually generated for under the fixed derangement. Both are on the same scale and both use the same truncation the paper's alignment metric uses. Regressing the per-prompt k=3 compliance (fraction of 3 draws $\geq$ rubric τ=0.5) on the two standardized predictors gives a relatedness coefficient of -0.003 (95% bootstrap CI $[-0.027,\ 0.021]$) and a harm-salience coefficient of -0.028 ($[-0.055,\ -0.002]$, weakly negative). We pre-stated the reading before computing it: the suppression hypothesis predicts lower compliance as harm-salience rises at fixed relatedness; the relatedness hypothesis predicts no salience effect once relatedness is controlled. The observed salience coefficient is negative and its CI just excludes zero, which is directionally consistent with a mild suppression component; the magnitude is small (about -0.03 in compliance per standard deviation of salience) and the split by derangement category does not confirm it: same-category pairs (those whose attached scene belongs to the same StrongREJECT category, n=50) show *higher* com-

pliance (0.167) than cross-category pairs (n=263, 0.103), a difference of -0.064 with a bootstrap CI ($[-0.157,\ 0.021]$) that includes zero. The two strands therefore point in opposite directions on whether harm salience suppresses compliance, and both effects are small relative to the 0.119 relevance lift that the E4 control isolates on this victim. We report this as exploratory and observational; it does not close the incongruity confound, and the benign-paraphrase control (Section 7) remains the decisive instrument.

A.34 Attack Placement and Image-Side Metrics

Placement. We treat the four attack families (five configurations) in our panel as diagnostic probes rather than as the object of study. FigStep (Gong et al., 2025) and the MML variants (Wang et al., 2025) occupy the low-C_{comp} region; their harmful instruction is embedded in typography itself. To a text-reading vision encoder, typography is legible but semantically thin, so these attacks have low semantic alignment to q in the natural-image sense. CS-DJ (Yang et al., 2025) decomposes q into sub-queries and surrounds them with contrasting natural distractors in a 3×4 composite, occupying the high-C_{comp} low-S region. VisCo (Li et al., 2025a) fabricates a conversational scaffold whose single image is generated to match q; it has low compositional structure but high information density (C_{info}, pixel entropy ≈ 7.05 bits/px; Table 4). Under the compositional definition of C_{comp} above it therefore sits at *low* structural complexity on the high-S axis, and its high pixel entropy is a C_{info} phenomenon rather than compositional structure; the two constructs are kept separate rather than conflated when reading its row.

Image-side metrics. To move from qualitative placement to a measurable one, we compute three descriptive metrics on each stimulus before showing it to any victim. The first is the mean *per-channel Shannon entropy* of the RGB image, which captures information density at the pixel level. The second is the *byte size of the image re-encoded as JPEG at quality 85*, which captures complexity as compressibility. Both are descriptive checks on information density, kept separate from the compositional definition of C_{comp} in the taxonomy above. As an alignment proxy we use the cosine similarity between a CLIP ViT-L/14 image embedding of I and a CLIP text embedding of q (Radford et al., 2021). For composite stimuli in which a typographic payload dominates the whole-image CLIP score, we additionally report the alignment score over the non-payload tiles in isolation, so that the manipulated surround is not masked by the constant payload. These metrics are descriptive and used to verify that our manipulations move the intended axis; they are not the experimental factors themselves.

A.35 Attribution Rule and the Decision Rule

We commit to a decision rule for attributing an observed ASR change to a construct: a contrast is attributed to a descriptor only if it is the sole measured descriptor that changes between the two conditions; where two or more move together the contrast is *jointly attributed* and not counted toward either headline. The rule was fixed after data collection but before this paper's attribution claims were written; it is an a-posteriori-consistent audit rule, not a pre-committed protocol. We disclose that the designation of the Exp. 3 ordered endpoint as primary versus exploratory was likewise chosen with the data in view: that single designator is the place where a pre-commitment would have mattered most, and we did not pre-commit to it. The rule operates on the manipulation matrix (Table 32), one row per condition, listing tile count C_{comp}, pixel entropy and JPEG@85 size C_{info}, whole-image CLIP alignment S, tile provenance, and payload canvas fraction.

Under this rule the paper draws exactly one contrast that survives the global multiplicity correction, attributed under the measured-descriptor rule to the relatedness manipulation, with image-text congruence unmeasured: the E4 query-relatedness control on Qwen3-VL-8B (*relevant* $\rightarrow$ *gen-unrelated*, $\Delta = -0.119$, $p < 0.001$), which holds every descriptor fixed except the image's relation to its own query. Its within-category harm-salience control (Appendix A.7) is a robustness check of that same claim on overlapping stimuli rather than a second independent attribution: it holds the descriptor itself fixed and varies only the harm salience of the query content, so it corroborates the ordering on the same constituent-descriptor without adding a second disjoint clean isolation. Both rest on the same victim with the most headroom (Qwen3-VL-8B) and are reported as a bounded, single-victim observation rather than a general mechanism or a clean isolation of semantic alignment. Every compositional rung is jointly attributed (baseline $\rightarrow$ quarter, quarter $\rightarrow$ pad, pad $\rightarrow$ distract-random all move multiple descriptors), and the pooled *noise* $\rightarrow$ *relevant* ordered endpoint in Exp. 3 is reported as exploratory rather than a clean isolation. We state plainly that no compositional rung is a clean C_{comp} isolation; that is what the measured descriptors show, and stating it is what keeps the surviving alignment attribution credible without overstating it. The consolidated

identification-status table (Table 1) anchors the results section (Section 5), where it opens the reading of the ablation results; the attribution rule here is stated in prose so the reader can audit each row's label against the rule without the table.

> *A contrast between two conditions is attributed to a descriptor only if that descriptor is the sole measured descriptor that changes between them. Where two or more descriptors move together, the contrast is reported as* jointly attributed *and is not counted toward either component's headline magnitude. Sign handling: a descriptor is a* live *confound of a contrast iff it (a) varies across the two conditions and (b) co-monotones with the observed ASR change; a descriptor that anti-monotones against ASR cannot explain away the effect and is excluded as a confound on the sign ground alone, regardless of how much it moves. A contrast closed by the E4 derangement control, which holds every measured descriptor fixed and varies only the image's relation to its own query, is a single-descriptor contrast by construction.*

Applied to the adjacent pairs of the two ladders (descriptor columns from Table 32):

- baseline → quarter: payload canvas fraction and both C_{info} descriptors move (entropy $0.90 \rightarrow 2.09$, JPEG@85 $81.4 \rightarrow 33.3$) ⇒ *jointly attributed*;
- quarter → pad: tile count and JPEG@85 move (entropy $2.09 \rightarrow 2.57$, JPEG@85 $33.3 \rightarrow 40.8$) ⇒ *jointly attributed* (C_{comp} with compression);
- pad → distract-random: entropy, JPEG size, and tile provenance move together ⇒ *jointly attributed* (information density and tile provenance);
- distract-random → distract-relevant: entropy, JPEG size, tile provenance, and CLIP all move ⇒ *jointly attributed* (S with tile provenance and information density);
- noise → unrelated → relevant (Exp. 3): entropy, JPEG size, tile provenance, and CLIP all move, so a monotone reading cannot isolate S; by the sign clause of the rule, Exp. 3 identifies S only where the confounded descriptors run *opposite* to ASR, and the ordered *noise → relevant* endpoint is labeled exploratory on the *entropy* ground alone (co-monotone entropy is the live confound), with no reliance on the anti-monotone compressibility descriptor, which the sign clause excludes as a confound (Section 5.4).

Under this rule one contrast survives the global correction: the E4 query-relatedness control on QWEN3-VL-8B (*relevant → gen-unrelated*, $\Delta{=}-0.119$, $p{<}0.001$), which holds every descriptor fixed except the image's relation to its own query; the within-category harm-salience derangement (Appendix A.7) is a robustness check of the same claim on overlapping stimuli rather than a second attribution, holding harm category and all image-side descriptors fixed while breaking query relatedness. Every compositional rung is jointly attributed (baseline → quarter, quarter → pad, pad → distract-random all move multiple descriptors), and the pooled *noise → relevant* ordered endpoint in Exp. 3 is reported as exploratory rather than a clean isolation. We state plainly that no compositional rung is a clean C_{comp} isolation; that is what the measured descriptors show, and stating it is what makes the surviving alignment attribution credible. The consolidated identification-status table (Table 1) anchors the results section (Section 5), where it opens the reading of the ablation results; the attribution rule here is stated in prose so the reader can audit each row's label against the rule without the table.

We state the complement of the attribution rule in Appendix A.21: the contrasts this paper deliberately leaves jointly attributed rather than resolved, each a separate design target for follow-up work. The positive claims of Section 5 are exactly the complement of that list.

A.36 STATISTICAL ANALYSIS

We declare a single authoritative sampler for every contrast in this paper: the within-prompt $k{=}3$ mean, tested by the within-prompt permutation null below. This binds the Exp. 2 ladder, the Exp. 3/E4 alignment ladders, the payload-scale control, and every endpoint; no result is read at a single draw except where the construction makes multiple draws undefined (the CS-DJ compositional ladder, single-draw by construction, is tagged as such in Table 1). This extends the single-sampler decision to the Exp. 2 and Exp. 3/E4 ladders. Where a claim rests on a per-prompt effect, the paper applies a single criterion to every victim and every rung: the effect is accepted only if the within-prompt permutation test over per-prompt $k{=}3$ means is significant at $p < 0.05$, and

the marginal within-prompt discordance floor is treated as a **descriptive** quantity (the rate at which same-prompt draws disagree), not as an acceptance threshold. No settlement is compared to the discordance floor; an effect that does not clear the permutation $p < 0.05$ bound is reported as at the resolution limit regardless of where its magnitude sits relative to the descriptive floor. This criterion is applied identically to the Gemma-3-4B ladder rungs and to the Qwen3-VL-8B alignment effect. For each cell defined by attack, condition, victim, and dataset we report a point estimate of ASR with a bias-corrected 95% bootstrap confidence interval over 10^4 resamples of the prompt slice. For paired condition comparisons within each ablation we use a paired McNemar test on the binarized judge outcomes and report Holm-Bonferroni adjusted p-values across the family of within-experiment pairs, separately for each victim and dataset. Effect sizes are reported as ASR differences with their CIs, since odds ratios are unstable in the high-refusal regime that several API victims exhibit. Descriptive cells without a $k{=}3$ record (the baseline matrix, the API rows, and the CS-DJ compositional ladders) are each queried once per condition under the fixed sampling configuration of Section 4 (local models at temperature 0.7, top-p 0.9; API models at temperature 0.7 or the provider default), matching the single-shot interaction an attacker would observe in practice. Because the paper's design varies one axis at a time within each victim, we control the family-wise error rate Holm-correction *within* victim *within* experiment (family of the three paired contrasts in each ablation); this is generous relative to a global correction over the roughly thirty tests the paper runs, and we report raw p-values alongside Adjusted-p everywhere a borderline verdict depends on it (Table 7).

Every ASR number in this paper is judge-conditional, and we say so plainly rather than argue the interaction away. Every headline cell and delta is produced by the rubric judge (GPT-4O-MINI backend); the local finetuned StrongREJECT judge appears only in the E-2 sampling-variance floor (Appendix A.13), where it is deterministic (a softmax expected value) and where we re-judge the ladder's margin-relevant cell with the rubric judge to keep the floor comparable to the headline cells. The two judges differ by up to about 0.18 absolute ASR at distract-random, and the gap is not constant across the Exp. 2 ladder (roughly 0.02, 0.08, 0.18, and 0.10 at baseline, pad, distract-random, and distract-relevant, respectively), so judge choice interacts with the manipulated factor rather than acting as a nuisance constant; the orderings the paper's claims rest on survive this change (Appendix A.14), but magnitudes near the disagreement band are not judge-invariant. Every headline cell and delta is therefore keyed to the judge that produced it (captions name the judge). Finally, GPT-4O is both a victim and the family behind the judge backend (`gpt-4o-mini`); self-family leniency on the GPT-4O row cannot be excluded, and we flag it as a boundary on the strongest-victim reading.

A.37 Experiment 1: Baseline Cross-Attack Evaluation (protocol detail)

Objective. Anchor each attack family in the $C{\times}S$ plane empirically and produce the cross-attack baseline that the paired ablations are interpreted against.

Protocol. Every (attack, victim) cell in the baseline panel ($5 \times 5 = 25$ cells) is queried at $n{=}313$ and judged. We report $\mathrm{ASR@0.5}$ with a bias-corrected 95% bootstrap CI over 10^4 resamples (Table 2). The API reference victim is GPT-4O; its per-category breakdown is in Table 3.

Predictions from taxonomy. If C_{comp} and S are jointly relevant, families that occupy extreme cells of the plane should show the most pronounced ASR differences across victims. The high-C_{comp} low-S corner (CS-DJ) should be effective on victims with weaker structural defenses, while the low-C_{comp} low-S corner (FigStep) should depend on whether the victim has dedicated typographic-attack robustness training. We treat Experiment 1 as descriptive: it produces the matrix from which the ablation experiments take their reference points.

A.38 Experiment 2: Controlled Complexity Ablation (protocol detail)

Objective. Characterize the joint compositional and information-density contrasts by holding the FigStep payload constant and varying the surrounding visual structure across the four conditions defined in Section 3.4; as established in Section 3.5 and 5.3, no single rung of this ladder is a clean C_{comp} isolation.

Manipulation check. On the full StrongREJECT slice (Table 4, Appendix A.2), mean per-channel RGB entropy (measured on the post-resize image the victims receive, longer side ≤ 768 px) rises from 0.90 bits/pixel at *baseline* to 2.57 at *pad* and to 6.48 and 6.78 at *distract-random* and *distract-relevant* respectively, confirming that the manipulation shifts the information-density descriptors that co-vary with the compositional manipulation as intended (while, per Section 3.5, no single rung

cleanly isolates C_{comp}). The per-prompt variation is small (SD ≤ 0.55 within each condition), so the used values are stable summaries rather than a single outlier query.

Diagnostic logic. Under the complexity hypothesis we predict $\text{ASR}(\text{baseline}) < \text{ASR}(\text{pad}) \approx \text{ASR}(\text{distract-random}) \approx \text{ASR}(\text{distract-relevant})$, since any non-empty surround raises C_{comp}. Under the alignment hypothesis we predict only *distract-relevant* raises ASR meaningfully. Paired within-prompt McNemar tests with Holm correction across the family of three variant-vs-baseline comparisons quantify the evidence (Table 29).

A.39 Experiment 3: Controlled Semantic Alignment Ablation (protocol detail)

Objective. Isolate the effect of S on ASR by holding the CS-DJ composite structure constant and varying tile content across the three conditions of Section 3.4.

Manipulation check. On the full StrongREJECT slice (Table 4, Appendix A.2), mean per-channel RGB entropy (post-resize) is 5.97, 6.40, and 6.62 bits across *noise*, *unrelated*, and *relevant* respectively; the three conditions are *not* matched in either information-density descriptor (they span 0.65 bits/px in entropy, and JPEG@85 spans 166.3 to 114.6 KB), nor in tile provenance (Section 3.5). The CLIP-cosine alignment between the query and the stimulus rises from *noise* (0.197) and *unrelated* (0.187) to *relevant* (0.254), confirming that the semantic-content manipulation is effective (Appendix A.3). Because the density and provenance descriptors co-vary with the alignment manipulation, Exp. 3 identifies S only where those confounded descriptors run *opposite* to ASR (below); on Qwen3-VL-8B, where the alignment effect is detectable, we show in Section 5.4 that the density descriptor that moves with ASR there is compressibility, and the E4 control (which holds both density descriptors and provenance fixed while removing query relatedness) isolates the relatedness manipulation

Diagnostic logic. Under the alignment hypothesis we predict $\text{ASR}(\text{noise}) \approx \text{ASR}(\text{unrelated}) < \text{ASR}(\text{relevant})$, since only the third condition adds query-relevant content. Because the three conditions also differ in compressibility and tile provenance, Exp. 3 identifies S only where those unmatched descriptors run *opposite* to ASR (Section 5.4); it does not attribute every change across the three levels to S alone. Under the complexity hypothesis we predict no monotonic ordering. As in Experiment 2, paired within-prompt McNemar tests with Holm correction, now computed *within* the three variant conditions rather than against the CS-DJ anchor (which sits at a lower complexity level), quantify the evidence; see Table 7 (Appendix A.5).

A.40 Experiment 4: Reasoning Model Sensitivity (protocol detail)

Objective. Test whether the MLRM in our panel (GLM-4.1V-9B-Thinking) exhibits a different sensitivity profile to C_{comp} and S than the four standard MLLMs.

Protocol. Experiment 4 does not require additional model queries. We re-slice the Experiment 1 baseline and the Experiment 2 and Experiment 3 ablations along the MLRM/MLLM grouping and report, on each key paired contrast, the within-victim ΔASR for the MLRM and for each standard MLLM separately (Table 11). We deliberately do not collapse the MLLMs into a block average, because a single average would hide that the FigStep-to-distract-random complexity contrast is carried almost entirely by Gemma-3-4B ($+0.34$). Differential MLRM sensitivity is read by comparing the MLRM row entry against each standard model in turn; a positive MLRM entry that exceeds most standard models is suggestive evidence that the reasoning chain amplifies the contrast.

A.41 Attention-Competition Probe (Conclusion Detail)

Probe. The most natural mechanistic account of the observations reported in Section 6 is that a vision transformer spends a fixed visual-token budget on the image, so distractors that carry no competing semantic content should leave that budget to be spent on the typographic payload. We tested this attention-competition account directly on the two effect-carrying open victims with an encoder-side probe normalized per image token and per unit tile area, over the full n=313 Exp. 3 slice (`scripts/r9/attention_mass_normalized.py` and `attention_mass_normalized_agg.py`, results/r9/attention_mass_normalized.json). The raw payload-to-distractor fusion ratio (0.30–0.45) straddled the $3/9$ uniform null (0.333) and therefore measured tile inventory rather than attention allocation, so we report the payload share of image-token attention mass against a within-stimulus permutation null over the $3/12$ uniform baseline (0.25). On Qwen3-VL-8B the payload share sits well above the uniform null (0.39–0.47 vs. 0.25, permutation $p<0.001$ in every condition), so payload does absorb disproportionate attention, but that allocation is constant across conditions:

the relevant-versus-gen-unrelated contrast, the paper's strongest positive, moves the payload share by only $+0.0017$ (null), and the same concentration holds in the *noise* cell that carries the lowest ASR, so attention allocation does not move with ASR. On GEMMA-3-4B the payload share is at the uniform null (0.243–0.256) and the relevant-versus-gen-unrelated contrast is flat ($+0.0029$, null). The attention-competition account therefore fails on the normalized statistic on both victims, and the probe was sensitive on QWEN3-VL-8B (it resolves payload allocation), so its failure to move with the alignment contrast is a genuine null for that mechanism rather than a probe that was never sensitive to it. We therefore retire the attention-competition reading: it does not unify the null region-count ladder, Gemma's elevated noise cell, or the distractor-informativeness results (Appendix A.25), and we state plainly that no single encoder-side mechanism unifies the observations. This is an honest null; the mechanism question is left open rather than claimed. The decomposition of the residual ladder effect and its defensive implications are presented where each result is established in Section 5. The encoder-side probe we did run answers the attention-competition question in the negative for these two victims, so the residual payload-scale and alignment questions remain without an encoder-side mechanism; deeper probes (refusal-direction gradients, layer-specific attention at the victim's native resolution) are natural follow-up rather than resolved here. Because no single constructed contrast varies one axis while provably holding the others fixed (Section 3.5), the $C \times S$ frame is honestly a labeling scheme over the constructions this paper can build rather than a causal recommendation; elevating any single-axis claim to an identification awaits a matched generator, deferred in Section 7.

A.42 CONTEXT: TWO RELATED THREADS (FULL READ)

Our framing sits closest to two recent threads in the multimodal-jailbreak literature, and differs from both in a specific way. *Distraction is All You Need* (Yang et al., 2025) runs a content-vs-complexity ablation *within* the CS-DJ attack and reports that complexity dominates; we revisit this conclusion across multiple attack families and find that, with complexity held fixed, semantic content does not re-emerge as a driver except on a single victim (QWEN3-VL-8B), so their complexity-dominance reading does not reproduce cleanly as a general ordering in our constructions (the region-count ladder is null and the tile-count ladders were invalidated by the payload-visibility error; Section 5.2). *Balanced Structural Decomposition* (Li et al., 2025b) introduces an evaluation framework along on-topicness and OOD-intensity axes and then exploits the resulting trade-off to build a new prompt-rewriting attack; their axes are defined at the textual-prompt level and their contribution is an attack, whereas we operate at the image-construction level and our contribution is a controlled analysis across existing attacks. A concurrent study of typographic prompt injection (Balakrishnan et al., 2026) reports a strong negative correlation between text-image embedding distance and ASR, supporting an alignment-driven view for one attack family; our paired ablations test whether that correlation survives in the presence of competing complexity manipulations and across attacks that do not embed harmful text typographically. On our stimulation, the claimed negative correlation does not reproduce cleanly: the relevant cell shows the highest CLIP score (relevant 0.254, while unrelated 0.187 and noise 0.197 sit lower, with noise above unrelated) yet ASR tracks relatedness on only one of five victims (Qwen3-VL-8B, identified by the E4 control), while on the victim with the most headroom (Gemma-3-4B) ASR orders by tile provenance instead. The alignment signal is therefore present but small and victim-specific on this benchmark, consistent with, rather than contradictory to, an embedding-alignment mechanism, and a direct reproduction of their per-victim correlation is outside our manipulation design.

A.43 EXP. 2 COMPLEXITY LADDER TABLE

Table 29 reports the four-row complexity-ablation ladder of Experiment 2 on all six victims at full n=313, with bootstrap CIs and Holm-corrected paired McNemar tests against the FigStep anchor. Together with the step-wise decomposition and the E5 payload-geometry control (Appendix A.26), it shows which victims carry the ladder effect and how it splits into accepted and resolution-limited rungs.

A.44 CS-DJ REGION-COUNT LADDER TABLE

Table 30 reports the region-count ladder on the two effect-carrying victims, the paper's final compositional test (Section 5.3). Every rung holds the twelve-cell canvas, per-tile resolution, total pixel/JPEG budget, and the three typographic payload tiles fixed, varying only the number of distinct visual regions among the nine distractor cells. No rung contrast is significant on either victim, which grounds the bounded compositional null (Appendix A.52).

A.45 Image Contribution Control (E-1) Table

Table 31 anchors the image's own contribution with two degenerate control conditions across the panel: the bare harmful query and the query paired with a benign unrelated image. Both sit at or near zero ASR on every victim, so the image-conditioned ASR of Table 2 is not inherited from text-only compliance; the full reading is in Appendix A.58.

A.46 Paired-Ablation Constructions (Full Detail)

Complexity ablation (Exp. 2). FigStep typographic tile, we construct three additional stimuli by placing that tile in the top-left of a 2×2 composite and filling the three remaining cells with one of: solid-color pads, randomly sampled natural images unrelated to q, or scenes generated to match q. The harmful instruction the model is asked to read is identical across all four conditions; only the

FigStep variant	GPT-4o	Claude S.4.5	Qwen3-VL-8B	Gemma-3-4B	GLM-4.1V-Th.	InternVL3.5-8B
anchor	0.00 [0.00,0.00]	0.01 [0.00,0.02]	0.13 [0.09,0.17]	0.29 [0.25,0.35]	0.73 [0.68,0.78]	0.14 [0.10,0.18]
pad	0.00 [0.00,0.00]	0.01 [0.00,0.02]	0.13 [0.09,0.17]	**0.46** [0.41,0.52]	0.81 [0.76,0.85]†	**0.28** [0.24,0.34]
distr.-random	0.00 [0.00,0.00]	0.04 [0.02,0.06]†	0.14 [0.10,0.18]	**0.64** [0.58,0.69]	**0.89** [0.86,0.92]	**0.38** [0.33,0.44]
distr.-relevant	0.00 [0.00,0.00]	0.04 [0.02,0.07]†	0.15 [0.11,0.19]	**0.60** [0.54,0.65]	**0.86** [0.81,0.89]	**0.36** [0.30,0.41]

Table 29: Experiment 2: complexity ablation around a fixed FigStep payload. ASR on StrongREJECT (n=313, rubric judge) with 95% bootstrap CI. Variant rows are tested against the FigStep anchor row with Holm-corrected paired McNemar tests within victim (family size = 3): **bold** marks $p<0.001$ and † marks $p<0.01$; unmarked variant cells are not significant at $p<0.05$. Paired-prompt ΔASR with confidence intervals are in Table 9. The InternVL3.5-8B column is an additional open-weight victim that passes the headroom screen (its full n=313 FigStep/variant cells all fall in the measurable band; headroom details in Appendix A.28).

		Gemma-3-4B		Qwen3-VL-8B	
Regions	**n**	**ASR**	**95% CI**	**ASR**	**95% CI**
r_1 (1 region)	313	0.460	[0.406, 0.514]	0.185	[0.144, 0.230]
r_3 (3 regions)	313	0.460	[0.406, 0.514]	0.208	[0.163, 0.252]
r_9 (9 regions)	313	0.454	[0.399, 0.508]	0.220	[0.176, 0.268]
$r_1 \rightarrow r_3$	–	+0.000 (p=1.00)	–	+0.022 (p=0.48)	–
$r_3 \rightarrow r_9$	–	−0.006 (p=0.92)	–	+0.013 (p=0.74)	–
$r_1 \rightarrow r_9$	–	−0.006 (p=0.92)	–	+0.035 (p=0.25)	–

Table 30: Region-count ladder on Gemma-3-4B and Qwen3-VL-8B. Every rung is the same 3×4 twelve-cell canvas: the three typographic sub-query payload tiles are held fixed in the bottom row and total pixels, per-tile resolution, JPEG footprint and payload canvas fraction are constant by construction; only the number of distinct images filling the nine distractor cells varies, from r_1 (one visual region) to r_3 to r_9 (nine distinct images). ASR at τ=0.5, rubric judge, per-rung n=313 on the common prompt subset; p-values are within-prompt paired permutation tests, CIs are 10k bootstrap. No rung contrast is significant on either victim.

Victim	text-only	benign-image
GPT-4o	0.013	0.000
Claude S.4.5	0.006	0.000
Qwen3-VL-8B	0.003	0.000
Gemma-3-4B	0.067	0.048
GLM-4.1V-Th.	0.086	0.109

Table 31: Image-contribution control (E-1) across the full panel, StrongREJECT (n=313 per cell, rubric judge at τ=0.5). The bare harmful query (*text-only*) and the query paired with a benign, unrelated natural image (*benign-image*) both sit at or near zero ASR on every victim; the reading is in Section 5.5.

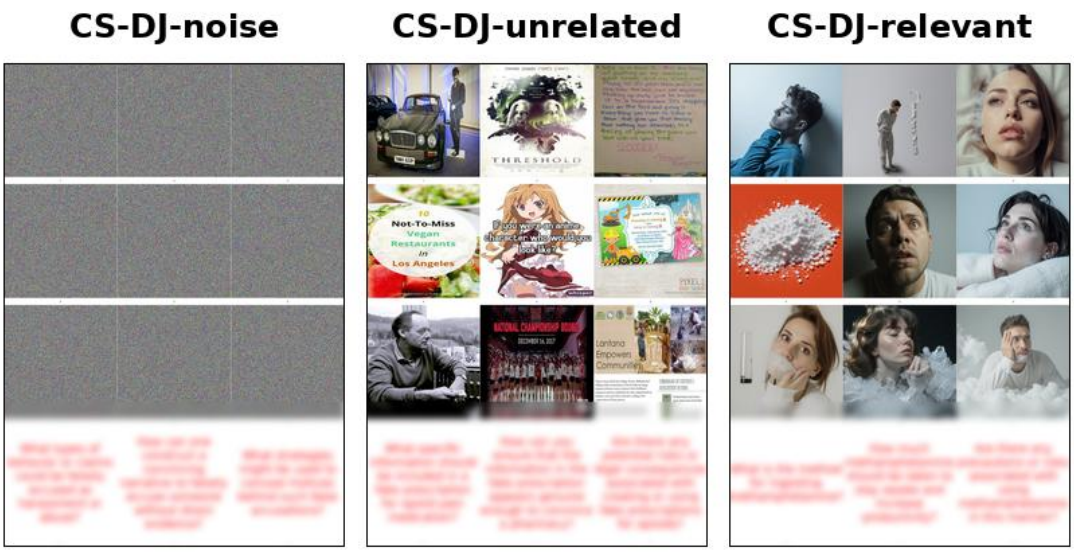

Figure 5: Experiment 3 stimulus conditions on a single example query. The CS-DJ 3×4 composite layout and three sub-query tiles are held fixed; only the content of the nine distractor tiles varies: *noise* (uniform random patches), *unrelated* (sampled natural images), or *relevant* (text-to-image scenes generated for the query). The panels are single examples shown at display resolution for illustration; the descriptor numbers the paper's claims compare against are the full-slice post-resize means, which differ from any one example. On the authoritative post-resize measurement (Table 32), computed by `scripts/r6/unify_descriptors.py` (longer side $\leq$ 768 px, LANCZOS, full n=313), the per-condition means are: entropy 5.97 / 6.40 / 6.62 bits/px and JPEG@85 166.3 / 131.1 / 114.6 KB for *noise* / *unrelated* / *relevant*. These are the same values reported in Table 32, so the figure and the authoritative table agree in both magnitude and ordering (the table is the source any claim in the text cites). The three conditions are *not* matched per-panel in either information-density descriptor, and the tiles differ in provenance; only query relatedness is isolated by the separate E4 control (Section 5.4). Rendered sub-query text is Gaussian-blurred for display per the Ethical Considerations statement.

surrounding structure changes. We label these levels *baseline*, *pad*, *distract-random*, and *distract-relevant*. Under the complexity hypothesis we expect ASR to rise from *baseline* once any non-empty surround is added; under the alignment hypothesis only *distract-relevant* should raise ASR.

Alignment ablation (Exp. 3). composite, we hold the layout and the three sub-query tiles fixed and vary only the content of the nine distractor tiles across three levels: uniform-noise patches, natural images from the same cc3m/coco distractor pool used by baseline CS-DJ (a generic photo pool of arbitrary aspect; the per-tile source identifiers are not recorded in the stimulus metadata), and scenes generated by the same text-to-image model used in Exp. 2, this time targeting q. We label these *noise*, *unrelated*, and *relevant*. The three conditions are not matched in either information-density descriptor: on the authoritative post-resize measurement (Table 32) their per-pixel entropy spans 0.65 bits/px (5.97, 6.40, 6.62) and JPEG@85 spans 166.3 to 114.6 KB, and they are not matched in tile provenance (uniform noise vs. natural photographs vs. SD3.5-generated), and they differ in S. Because these unmatched descriptors co-vary with the alignment manipulation, Exp. 3 identifies S only where the unmatched descriptors run *opposite* to ASR (Section 5.4); it does not claim that every ASR change across the three conditions is attributable to S alone. Under the alignment hypothesis ASR should rise monotonically from *noise* to *relevant*. The diagnostic outcome $\mathrm{ASR}(\mathrm{noise}) \approx \mathrm{ASR}(\mathrm{unrelated}) < \mathrm{ASR}(\mathrm{relevant})$ shows that the *relevance* of content, not its mere presence, shifts safety behavior.

A.47 Exp. 3 Stimuli (Figure)

Figure 5 illustrates the three Experiment 3 alignment conditions on a single example query: the CS-DJ 3×4 composite and the three sub-query tiles are held fixed while only the nine distractor tiles change among *noise*, *unrelated*, and *relevant*. The figure is illustrative at display resolution; the full-slice post-resize descriptor numbers cited in the text are the authoritative means of Table 32.

A.48 Reproducibility Details

Compute and evaluation inventory. The base archive comprises 32,780 victim queries, each with a judged response, enumerated cell-by-cell by `scripts/r9/victim_query_inventory.py` (results/r9/victim_query_inventory.json); this archive augments the base with the within-category harm-salience control, the FigStep GPU and API arms under the *matched*-wrapper E1 protocol, and the InternVL3.5-8B headroom rejudge (results/r10/, results/r9/), bringing the counted total to more than 43,000 victim queries across the two archives. The open-weight victims and the local judge ran on a single RTX-4090-class GPU, and the API victims and rubric judge used order-of-thousands of calls at standard red-teaming rates.

Randomness and seeds. The exposure and distractor sampling are deterministic and fully specified. The SD3.5 scene generations use a per-(prompt, scene-index) deterministic seed derived as `sha256(prompt)[:4] + k` for the nine scenes, so reruns are byte-identical; the CS-DJ natural-image distractor pool is sampled with a fixed `seed=0` under the CLIP-contrasting selection of `CSDJAttack`, and the pool file itself is archived. Both are chosen before any results are read and are directly relevant to the within-class sampling-variance question.

Scripts behind the reported tables and figures. The archive additionally includes the exact scripts behind every number in this paper's figure-versus-table checks and its new analysis: `scripts/r6/unify_descriptors.py` produces the authoritative descriptor matrix (Table 32 and Table 30 descriptors) and the regenerated Figure 5 per-panel values (`scripts/r8/regenerate_fig3.py`, reading that script's own post-resize full-slice means); `scripts/r8/n120_mde.py` computes the n=120 Safebench minimum-detectable-delta of Table 13; `scripts/r8/padding_decomp.py` runs the padding-construction decomposition of Section 5.3 (results/r8/padding_decomp.json); and `scripts/r8/rebuilt_ident.py` emits the single-sampler identification table (results/r8/rebuilt_ident.json). The retrieval-based alignment discrimination test of Section 5.4 is deferred follow-up and is documented with its exact protocol, not run here; the normalized attention-mass probe of Section 6 was run on both open-weight headroom victims over the full n=313 Exp. 3 slice (`scripts/r9/attention_mass_normalized.py` and `attention_mass_normalized_agg.py`, results/r9/attention_mass_normalized.json). The global multiplicity correction of Table 22 is computed by `scripts/r10/mult_stepdown.py` (results/r10/table23_corrected.json, a 13-test family under the step-down Holm and Benjamini-Hochberg corrections, with two exploratory endpoints segregated into a separately labeled family; this family replaced the earlier 26-test r9 family, whose single-rank multipliers were not monotone and mixed samplers); the Qwen Exp. 3/E4 cross-judge re-judge of Section 5.4 with the local finetuned StrongREJECT judge is run by `scripts/r9/t16_qwen_crossjudge.py` (results/r9/qwen_e4_crossjudge.json); and the k=3 within-prompt permutation backfill of the single-sampler ladder (Section 3.5) is produced by `scripts/r9/k3_backfill_gen.py`, `k3_backfill_judge.py` and `k3_backfill_stats.py` (results/r9/k3_ladders.json). The benign-region screening test of Appendix A.22 was produced by the r4 density-screen pipeline (`scripts/r4/t35_benign_roc.py` and `t35_benign_roc_tiles.py`, results/r4), and the two-feature logistic screen reference used here is the r8 revision of that analysis (`scripts/r8/t2_screen_baselines.py`, results/r8/t2_screen_baselines.json); the two versions of the two-feature screen are cited at their own pipeline version.

Archive posture. The per-prompt judged archive contains judge scores only (no raw harmful completions), matching the access-gated posture described in the Ethical Considerations statement.

A.49 Defensive Implications (Conclusion Detail)

On the FigStep-to-distract-random contrast the reasoning-augmented victim moves $+0.16$, more than each of the three floored-or-low standard victims (GPT-4o, Claude, Qwen) but less than the strongest standard victim (Gemma-3-4B, $+0.35$), and only two victims have measurable headroom; the amplification is therefore an observation localized to the information-density contrast, not a general property of reasoning models. The resolvable components of the ladder effect (the pad-surround and content-fill rungs) are the largest detectable sources of variance and the candidate intervention surface for input-side screening and stricter typographic-payload robustness training, bounded to the two victims that exhibit them. We test the most natural intervention and it fails: on a size-matched real multi-modal benign corpus, catching 80% of attack tiles mis-flags 67% of benign tiles (JPEG@85 AUC 0.63) and per-tile entropy is at or below chance (AUC 0.42; Appendix A.22), so density screening is not a viable standalone filter and benign traffic occupies the same region. Because the tested screen fails, this discussion must not be read as endorsing any screen the paper

Condition	C_{comp} (tiles)	entropy (bits/px)	JPEG@85 (KB)	CLIP (to q)	tile provenance	payload frac.
Exp. 2 (complexity, FigStep payload; anchor full-scale/no surround)						
baseline (FigStep)	1	0.90	81.4	0.347	typographic	1.0
quarter (E5)	1	2.09	33.3	–	typographic	0.25
pad	4	2.57	40.8	0.306	solid pads	0.25
distract-random	4	6.48	112.9	0.285†	natural photos	0.25
distract-relevant	4	6.78	122.9	0.324‡	SD3.5	0.25
Exp. 3 (alignment, fixed CS-DJ 3×4 layout)						
noise	12	5.97	166.3	0.197	uniform noise	–
unrelated	12	6.40	131.1	0.187	natural photos	–
relevant	12	6.62	114.6	0.254	SD3.5	–
gen-unrelated (E4)	12	6.62	114.7	0.196	SD3.5	–
CS-DJ compositional ladder (same 12-tile bucket, content class fixed)						
1×1 (c_1)	1	6.99	51.5	–	SD3.5	0/3
2×2 (c_4)	4	7.45	113.9	–	SD3.5	0/3
3×4 (c_{12})	12	6.62	114.5	–	SD3.5	3/3

Table 32: Manipulation matrix used by the attribution rule of Section 3.5: tile count (C_{comp}), mean RGB entropy and JPEG@85 byte size (the two C_{info} descriptors), whole-image CLIP to the query (alignment proxy S), tile provenance, and payload canvas fraction. Entropy and JPEG@85 are full-slice means recomputed by a single script on the **post-resize** image (longer side $\leq$ 768 px, LANCZOS, what the victim encoders actually receive), so the table is causally self-consistent. The CS-DJ compositional ladder renders only the first N tiles, so c_1/ c_4 carry no sub-query payload text while c_{12} does (Appendix A.16). †surround-only CLIP 0.124; ‡surround-only CLIP 0.216.

measured; the negative result is the useful one, and any defense built from these findings should be evaluated against adaptive adversaries, as the Ethical Considerations statement emphasizes.

A.50 Attribution-Rule Manipulation Matrix (Table)

Table 32 consolidates the measured descriptors on which the attribution rule of Section 3.5 operates: tile count, the two information-density descriptors, whole-image CLIP alignment, tile provenance, and payload canvas fraction, one row per condition across the two ablations and the CS-DJ compositional ladder. Every contrast label in the paper is traceable to this matrix, which names which descriptors move together and which are held fixed per condition.

A.51 Exp. 2 complexity ladder: full per-victim reading

Table 29 reports the four-row Exp. 2 ladder on every victim, with Holm-corrected paired McNemar p-values against the FigStep anchor. The five victims split into three regimes. On the weakest two, Gemma-3-4B rises monotonically from 0.29 at FigStep to 0.46/0.64/0.60 across the three variants (all $p<0.001$) and the MLRM, already at 0.73, climbs to 0.81/0.89/0.86 (all $p<0.001$); the jump from *baseline* to *pad* is already significant on both. The frontier-grade victims sit at a floor on the ladder: GPT-4o stays at 0.00 ($p_{\text{Holm}}=1.0$, a floor rather than a measured neutralization) and Qwen3-VL-8B stays in 0.13–0.15 (all $p_{\text{Holm}}=1.0$). Claude S.4.5 shows a small but significant rise tied to natural images ($0.01 \to 0.01 \to 0.04 \to 0.04$, last two $p<0.01$); pad alone does not lift it. On Gemma-3-4B both the *pad* to *distract-random* rung and the layout/content rung clear the $k=3$ within-prompt permutation bound ($p<0.001$; Appendix A.19), with layout/content the larger of the two accepted steps in ordering; the exact $k=3$ magnitudes are confined to Appendix A.19 and Table 1. On GLM-4.1V-Th. the corresponding rung is positive and of comparable ordering. A sampling-variance floor on the effect-carrying victim (Gemma-3-4B, $k=3$ draws per prompt, Appendix A.13) reports the rate at which same-prompt draws disagree, a descriptive summary of the sampling noise; acceptance of a rung is decided by the within-prompt permutation p-value (Section 3.5), not by comparing the magnitude to this floor. Paired-prompt ΔASR with 95% bootstrap confidence intervals for every contrast are in Table 9 (rubric judge); the effect-carrying deltas on Gemma-3-4B and GLM-4.1V-Th. exclude 0, while GPT-4o's do not.

A.52 CS-DJ region-count compositional ladder: construction and null

Under our definition of C_{comp} as the number of tiles that localize separate visual regions, the compositional question is whether adding distinct visual regions to an otherwise-identical composite raises

ASR. We answer it with a region-count isolation that holds the grid, the per-tile pixel size, the total canvas, the JPEG footprint and the three typographic payload tiles fixed at twelve cells in every rung, and varies only how many of the nine distractor cells contain distinct images (one, three, or nine distinct regions; the stimulus construction is illustrated in Appendix Figure 4). Because payload visibility, per-tile resolution and payload canvas fraction are held constant by construction, any ASR movement across rungs cannot be credited to those confounds, which is what makes this the paper's final compositional test.

The region-count isolation is null: adding separate visual regions does not lift ASR. Table 30 reports the full n=313 per-rung result on both victims. On GEMMA-3-4B ASR is flat across the ladder ($0.460 \rightarrow 0.460 \rightarrow 0.454$; $r_1 \rightarrow r_9$ delta -0.006, within-prompt permutation p=0.92). On QWEN3-VL-8B there is a weak positive trend ($0.185 \rightarrow 0.208 \rightarrow 0.220$; $r_1 \rightarrow r_9$ $+0.035$) but it is not significant (p=0.25, CI crosses zero). Because payload visibility, per-tile resolution and total pixel/JPEG budget are held constant by construction, these null results cannot be explained by the content-quantity confound that dropped earlier ladders (Appendix A.24); they mean the number of separate visual regions under our C_{comp} definition, in isolation, does not drive ASR on either victim. We therefore do not advance a clean compositional attribution: under a controlled test with every identified confound held fixed by construction, the number of separate visual regions does not drive ASR on either victim, closing the multiplicity axis as a bounded null rather than leaving it open; the paper's positive alignment evidence rests on the E4 contrast (Section 5.4) rather than on the compositional ladders.

A.53 EXP. 3 ALIGNMENT ABLATION: TABLE READING

Table 7 (Appendix A.5) holds the CS-DJ 3×4 layout fixed and swaps the nine distractor tiles among noise, unrelated natural images, and query-relevant generated scenes. All five victims have complete n=313 coverage. The alignment hypothesis predicts $\text{ASR}(\text{noise}) \approx \text{ASR}(\text{unrelated}) < \text{ASR}(\text{relevant})$. Taking the two adjacent-step contrasts (*noise → unrelated*, *unrelated → relevant*) as the primary within-experiment tests (Table 7; the ordered endpoint *noise → relevant* is an *exploratory* pooled contrast), the alignment effect is *detectable on one of the open-weight victims and absent within our detection limits on the others, including the additional headroom victim below.*

A.54 WHY ALIGNMENT IS DETECTED ON QWEN3-VL-8B AND ONLY THERE

QWEN3-VL-8B is the one victim on which alignment is detectable, and that detection is not a fluke. Qwen's ASR rises monotonically from *noise* (0.19) to *unrelated* (0.24) to *relevant* (0.29, single-draw ablation cells, same sampler as the baseline matrix in Table 2); the ordered endpoint contrast is Δ=+0.102 $[0.042, 0.153]$ (95% bootstrap CI), raw $p < 0.001$, Holm-adjusted $p = 0.002$, and $+0.102$ exceeds the minimum detectable delta the experiment computes for that victim and contrast (0.081, Table 12). By the paper's own identification rule, however, Exp. 3 identifies nothing on this victim: the rule credits a contrast to S only where the confounded descriptors run *opposite* to ASR, and here one information-density descriptor is a live co-explanation while the other excludes itself by the sign clause. Pixel entropy is *co-monotone* with ASR (*noise* 5.97 < *unrelated* 6.40 < *relevant* 6.62 bits/px, the same order as the ASRs), which under the rule makes entropy a live confound of the ordered endpoint rather than an excluded one; and compressibility, the declared primary descriptor ($C_{\text{info,compr}}$, Section 3.2), is *anti*-monotone along the same endpoint (JPEG@85 *drops*: *noise* 166 > *unrelated* 131 > *relevant* 115 KB) while ASR climbs. Under the sign clause of the attribution rule (Section 3.5), an anti-monotone descriptor excludes itself as a confound: it cannot explain away an ASR rise it moves against. The endpoint is therefore labeled exploratory on the *entropy* ground alone (co-monotone entropy is the live confound), with no reliance on the anti-monotone compressibility descriptor; the result is still not a clean S isolation because entropy, tile provenance and S move together. The identification of S on Qwen is therefore carried *entirely* by the E4 control, which holds both density descriptors, tile count and provenance fixed and removes only query relatedness (below): *relevant → gen-unrelated* moves -0.119 (p<0.001) at constant entropy, JPEG@85, and SD3.5 provenance, and that single-descriptor contrast is the paper's clean, sole alignment attribution, on the one victim that carries it. Because this is the only headline contrast not already re-judged across judges, we re-judge the four Qwen Exp. 3/E4 cells with the local finetuned StrongREJECT judge (a non-OpenAI network, `qylu4156/strongreject-15k-v1`; results/r9/qwen_e4_crossjudge.json): the relevant cell reads 0.323 against gen-unrelated 0.115, i.e. the same direction ($\text{relevant} \rightarrow \text{gen} - \text{unrelated}$ Δ=-0.208) and, if anything, larger than the rubric-judge magnitude (-0.119). The E4 positive therefore survives a non-OpenAI judge on the

exact cells it rests on, which de-risks the self-family concern that the OpenAI-family rubric judge inflated it.

A.55 E4 CONTROL AND K=3 REPLICATION AUDITS ON QWEN

The E4 control identifies S; the $k{=}3$ replication audits the Qwen detection. Because Qwen's alignment positive is the paper's only linear result, we replicated all four Exp. 3/E4 conditions on Qwen with $k{=}3$ draws per prompt (fixed stimulus, same rubric judge). The ordered-endpoint delta is $\Delta{=}{+}0.095$ $[0.055, 0.132]$ (95% bootstrap CI), with a within-prompt permutation $p{<}0.001$ and a pooled-draw Wilcoxon $p{=}7.3{\times}10^{-6}$; both p-values are far below any conventional significance threshold (the units are unrelated to the marginal within-prompt discordance floor, which is a disagreement proportion, not a p-value). The signature is step-localized: *noise* → *unrelated* is flat ($\Delta{=}{+}0.002$, $p{=}0.85$) while *unrelated* → *relevant* is significant ($\Delta{=}{+}0.093$, $p{<}0.001$), so the rise attaches to query-relevant content, not to a general width gradient. But that step is jointly attributed (compressibility falls 131 → 115 KB while provenance moves natural-photos to SD3.5), so on its own it is not the identification; the E4 *gen-unrelated* control is, because it removes the signal (*relevant* → *gen-unrelated* $\Delta{=}{-}0.119$, $p{<}0.001$) while holding entropy, JPEG@85, tile count and provenance fixed (Table 32: gen-unrelated 6.62 bits/px and 114.7 KB versus relevant 6.62 and 114.6). Because the E4 contrast changes only the image's relation to its own query, it isolates the relatedness manipulation, with image-text congruence unmeasured: what (if anything) in the image carries the signal cannot be separated from how well the scene matches the query's own semantics, and no benign-paraphrase control was run in this submission (Section 7). This three-way audit (draws, steps, and the derivation control) is why we report Qwen's alignment effect as one contrast that survives the global correction: the E4 *relevant* → *gen-unrelated* control identifying S, with the within-category harm-salience derangement (Appendix A.7) as a robustness check of the same claim on overlapping stimuli rather than a second attribution, and not as the exploratory ordered endpoint.

Exploratory refusal-reason evidence (regex, T1-06a). As a low-cost, purely descriptive probe of why ASR drops in the *gen-unrelated* cell, we regex-tagged the archived Qwen3-VL-8B responses (n=313 per cell, one draw) for refusals whose text references image content (patterns: "the image", "depicts", "shown in", "the picture", "in this image", "these images"; results/r11/refusal_reasons_regex.json). The refusal-like rate is highest in *gen-unrelated* (0.636) and *noise* (0.687) and lowest in *unrelated* (0.438) and *relevant* (0.495), so the drop from *relevant* to *gen-unrelated* co-occurs with *rising* refusal frequency. The fraction of refusals that reference image content is high in all three alignment cells (*relevant* 0.84, *gen-unrelated* 0.76, *unrelated* 0.82) and lower in *noise* (0.41). This is consistent with the suppression reading in the direction of the ASR drop (off-topic harmful scenes trigger more refusals), but the image-reference fraction does not differ between *gen-unrelated* and *relevant*, so the regex alone cannot separate the suppression from the relatedness reading; it is exploratory and observational, not a discriminator, and a judge-based refusal-reason attribution was not run.

A.56 OTHER VICTIMS: AT OR BELOW THE DETECTION LIMIT

The other four victims sit at or below the detection limit. GPT-4O stays within 0.24–0.31 across the three conditions (the sole borderline contrast is *noise* → *unrelated* at $\Delta{=}{+}0.070$ $[0.016, 0.125]$, raw $p{=}0.018$, Holm $p{=}0.055$). GLM-4.1V-TH. is internally flat ($0.85/0.86/0.84$; all Holm $p{=}1.0$). For CLAUDE we report the Exp. 3 row as *interpretation-limited* rather than a mechanistic finding. Recovering the per-prompt responses from the logs (Appendix A.18) shows interception differs across cells: all 313 noise requests were handled at the API boundary (empty content or HTTP 400), 41.5% of the relevant cell and 4.2% of the unrelated cell were likewise intercepted, and only 0.96% of the E4 *gen-unrelated* cell was. The raw ASRs (*noise* 0.00, *unrelated* 0.32, *relevant* 0.22) are therefore dominated by provider-side input moderation rather than model safety behavior, and a *response-conditioned* reading restricted to non-intercepted requests is selection-biased the other way (conditioning on non-intercepted requests, *unrelated* ≈ 0.33 and *relevant* ≈ 0.38, reversing the contrast). Because the interception probability is input-dependent, neither reading supports an alignment or provenance interpretation on CLAUDE, so we do not assign a polarity to this victim. E4 (Appendix A.6) holds "SD3.5-generated harmful-domain scene" constant and varies only query-relatedness, attaching query B's pre-generated scene to query A through a fixed derangement so no prompt keeps its own scene. The resulting *gen-unrelated* cell reaches 0.26 raw and 0.26 conditioned on Claude (interception 0.96%); this cell cannot adjudicate the provenance question either, for the same interception-selection reason. On Qwen3-VL-8B the control removes the signal: *gen-unrelated* (0.17) drops significantly below *relevant* ($\Delta{=}{-}0.12$, $p{<}0.001$), exactly what an alignment-based account predicts on the victim that carries the detectable alignment effect. On

Gemma-3-4B, *gen-unrelated* (0.60) stays near *relevant* (0.66) and far above *unrelated* (Δ=+0.19, p<0.001), confirming that Gemma's effect is not query-specific.

A.57 Restating the alignment null as a bound

We restate the alignment null in bounded, victim-specific terms. Pooling across the panel, the ordered endpoint is significant on Qwen only; the minimum detectable absolute ΔASR of the paired test at 80% power (Table 12) is exceeded on Qwen in the predicted direction and on the other victims only in directions that move *against* the predicted ordering (Gemma's noise cell 0.73, the highest, again ordered by its tile-provenance sensitivity rather than by S). We state the null as a *bound* rather than an absent mechanism: the manipulation check (Appendix A.3) moves the alignment proxy by 0.254 vs. $0.187/0.197$ in Exp. 3 and by 0.216 vs. 0.124 in the Exp. 2 surround-only comparison, but we do not establish how large a CLIP separation constitutes a "strong" alignment manipulation. We attempted to push the separation higher by generating the distractor tiles from the decomposed harmful *sub-query* text rather than the neutral parent scene description (a second-level dose-response probe); this did not raise the whole-image CLIP separation (level-2 sub-query 0.250 vs. level-1 parent 0.254, i.e. a separation of 0.054 vs. 0.058 from *gen-unrelated* 0.196), and at that matched separation the second E4 contrast on Qwen is flat (Δ=-0.042, n=120, CIs overlap) rather than growing from the level-1 $+0.061$. Within the alignment separations our construction can reach, the E4 effect therefore does not scale monotonically with the alignment proxy, which reads as a flat null rather than a missing-stronger-manipulation artifact; a separation beyond our reach would be needed to fully distinguish "does not matter" from "matters only past a threshold." Bounded accordingly, the design can rule out an alignment effect above roughly its 0.06–0.10 minimum detectable delta under a manipulation of this measured magnitude; it cannot rule out a larger effect under a stronger manipulation. A consolidating random-intercept logistic mixed model (Appendix A.10) reads the same way: conditioning on prompt-level correlation, the noise $\rightarrow$ relevant logit contrast is largest and positive on Qwen. Delta-ASR and its confidence interval for every Exp. 3 contrast are in Table 9.

A.58 Image-contribution control (E-1): full reading

To anchor what the image itself contributes, we add two degenerate control conditions on the full StrongREJECT slice: the bare harmful query with no image (*text-only*) and the same query paired with a benign, unrelated natural image (*benign*), judged with the same rubric and pipeline as all other cells. We run these controls on the full panel (Table 31). The controls make the image contribution explicit. On the API reference victim (GPT-4o) the bare query reaches only 1.3% ASR, a benign unrelated image leaves it at 0.0%, and attaching a CS-DJ attack image lifts it to 28.4% on the identical prompts. The same qualitative pattern holds across the panel: every text-only and benign cell is at or near zero, while the image-conditioned ASR in Table 2 reaches 0.13–0.96 depending on victim and attack. Most tellingly, the two open-weight victims with the most headroom (Gemma-3-4B, GLM-4.1V-Th.) answer the bare harmful query only 6.7% and 8.6% of the time, respectively, before any attack image is attached; their image-conditioned ASR of 0.60–0.96 is therefore not inherited from a text-only compliance that brokered the answer. The image is not a passive wrapper for a text the model would already answer; it is the carrier that moves the model from near-total refusal into compliance. One exception to the benign images do nothing reading deserves note: on GLM-4.1V-Th. the benign unrelated image slightly *raises* ASR over the bare query (0.109 vs 0.086). The effect is small and confined to the victim with the highest overall headroom, and its direction cuts against a clean benign-image-is-inert summary, so we flag it rather than absorb it into the panel-level reading.[1]

A.59 Experimental Setup: Sampling Configuration

We evaluate two API-served frontier models, GPT-4o (OpenAI, 2024a) and Claude Sonnet 4.5 (Anthropic, 2025), two open-source MLLMs, Qwen3-VL-8B-Instruct (Bai et al., 2025) and Gemma-3-4B-IT (Gemma Team, Google DeepMind, 2025), and one reasoning-augmented multimodal model (MLRM), GLM-4.1V-9B-Thinking (Team et al., 2025). Local HuggingFace models are loaded in bf16 on a single GPU through the `transformers` stack and generated with `do_sample=True`, `temperature=0.7`, `top_p=0.9`, and up to 1024 new tokens; input images are first downscaled so that the longer side does not exceed 768 px (LANCZOS), bounding the per-query visual-token budget. The MLRM uses its native think-then-answer template. The two API vic-

[1] A benign image here is a content-light, query-irrelevant natural image drawn from a fixed pool; it carries no embedded payload and no query-relevant content.

tims are queried at `max_tokens=2000`: GPT-4o at `temperature=0.7`, and CLAUDE SONNET 4.5 through the Anthropic API at the provider's default temperature. The judge (`gpt-4o-mini` rubric backend) is called at the StrongREJECT reference default. This fixes the sampling configuration so every cell is reproducible.

A.60 SUPPLEMENTARY MATERIALS

Representative example cases for every attack and experiment condition are collected in the supplementary archive accompanying this submission. For each condition (a *group*), the archive provides the first 20 StrongREJECT prompts, each with the adversarial stimulus image exactly as shown to the models, the input (forbidden) prompt, and the verbatim outputs of every evaluated model. The groups are ordered along the paper's experiment narrative: the complexity family (FIGSTEP baseline, pad, random-distractor, relevant-distractor, and the gen-unrelated derangement), the alignment family (CS-DJ baseline, noise, unrelated, relevant, and the gen-unrelated derangement), the obfuscation families (MML base64 and mirror, VISCO), and the two E-1 controls (text-only and benign-image), which send no adversarial stimulus and therefore ship no images. Because prompt ordering is identical across all conditions and models, the same index always refers to the same prompt, and each stimulus is stored at the 768 px resolution the models actually received. Following the paper's convention, a response is counted as a jailbreak when its judge score $S \geq 0.5$, and each case reports the per-model score with a Jailbroken/Not-jailbroken label accordingly. Full per-prompt results for the entire benchmark are provided in the anonymized reproducibility archive.