

Hard to Read Is Not Hard to Audit: Why Degrading a Text Image Raises Jailbreak Success in One 2026 Study and Lowers It in Another, Four Accounts That Fit the Same Curve, and the Ablations That Would Separate Them

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Rendering text as an image is now proposed as a way to compress long contexts for multimodal language models, and a 2026 study reports that the same move opens a safety hole: when a harmful request is rendered at a resolution low enough to be hard but not impossible to read, attack success rises sharply and then falls again as the image sharpens. Its authors explain this as cognitive overload - the effort of reading diverts the model from checking - and propose a prompt that makes the model transcribe first and audit second. A second 2026 study, which tested one of the same models, finds the opposite sign: smaller fonts and heavier blur lower attack success, and for two frontier models clean renders are less dangerous than plain text. In this analysis the author reviews and synthesises published results to show that the two studies are not in contradiction once the prompt framing, the dose of degradation and the scoring rule are held apart, and that the inverted-U curve the first study reports is consistent with at least four accounts: resource contention, delayed emergence of harmful features in depth, and a gap in what safety training covered each produce it, and the way success is scored changes its height. Re-reading the first study's own per-model sweeps, the author shows that seven of its ten models show the inverted U, one rises without returning, two are flat, and three of them refuse clean renders more often than plain text. None of the study's ablations, including its defence, discriminates among the four accounts. The paper states the predictions on which the accounts differ, gives a decision procedure for what a degradation result licenses, and specifies five measurements that would settle which account holds.

Keywords: multimodal large language models; jailbreak; typographic attacks; visual degradation; safety alignment; mismatched generalization; visual-text compression; attack success rate

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

Multimodal large language models (MLLMs) read text in images well enough that rendering text as pixels has become a proposal for saving tokens. DeepSeek-OCR reports about 97 percent decoding precision when the number of text tokens is under ten times the number of vision tokens (Wei, Sun and Li, 2025). Glyph renders long inputs into images and processes them with a vision-language model to extend the usable context (Cheng et al., 2025), and a study of rendered inputs to general-purpose MLLMs reports token savings that are often close to half on long-context retrieval and summarisation without loss of task accuracy (Li, Lan and Zhou, 2025). Pixel-based language modelling predates all three (Rust et al., 2022). The appeal is that a model which reads pixels need not care how a string is tokenised.

The safety literature has known for some time that putting a request into an image changes how a model treats it. Typographic renders of prohibited instructions reach an average attack success rate (ASR) of 82.5 percent across six open LVLMS (Gong et al., 2023); query-relevant images compromise MLLMs whose language backbones are safety-aligned (Liu et al.,

2023a); crafted images lift ASR to 90.26 percent on LLaVA-1.5 and 71.60 percent on Gemini Pro Vision (Li et al., 2024a); and adversarially optimised images can override a model's behaviour outright (Qi et al., 2023; Bailey et al., 2023; Shayegani, Dong and Abu-Ghazaleh, 2023). Misleading text placed in an image also degrades LVLMS classification, by up to 60 percent when the attack text is suggested by an LVLMS itself (Qraitem et al., 2024), and jailbreak prompts written for text-only models transfer to MLLMs with high success on ten open models (Luo et al., 2024), which locates part of the multimodal vulnerability in the text-processing path rather than in vision. The usual diagnosis is that safety behaviour learned in text does not carry over to the visual pathway: hidden states for images sit away from those for equivalent text at the layers where the safety mechanism acts (Xu et al., 2024), multimodal representations drift from the text distribution the backbone was aligned on (Liu et al., 2024b), and vision-language fine-tuning can erode the alignment the backbone had (Zong et al., 2024).

Song et al. (2026) add a new variable to this picture: the legibility of the rendered request. They render 770 role-play jailbreak instructions at nine resolutions from 15 to 300 DPI, query ten MLLMs, and report that ASR follows an inverted U - near zero when the text is unreadable, highest in an intermediate band where the model can

still transcribe most of the text, and lower again when the image is sharp. They call the band the Attack Comfort Zone (ACZ), attribute it to "Cognitive Overload", and propose Structured Cognitive Offloading (SCO), a prompt that asks the model to transcribe the text, then evaluate it for harm, then respond. On Qwen3-VL the defence takes ASR from about 67.4 percent to 4 percent. If the effect is general, it matters beyond red-teaming: a deployment that downsamples rendered context to save tokens would be choosing, as an engineering parameter, the resolution at which its safety behaviour is weakest.

Balakrishnan, Mendapara and Garg (2026) ran an experiment that overlaps this one closely and reached a different answer. They render 1,000 SALAD-Bench attack prompts at twelve font sizes from 6 to 28 pixels and under ten visual transformations, query four VLMs including Claude Sonnet 4.5, which Song et al. also tested, and find that ASR is near zero at 6 pixels, rises to a plateau by 10 to 12 pixels, and is driven down - not up - by heavy blur. For GPT-4o and Claude Sonnet 4.5, clean renders are less effective than the same prompt as plain text: 35.6 percent text against at most 7.7 percent image for GPT-4o, and 46.6 against at most 21.6 percent for Claude. Best-of-N jailbreaking, which randomly perturbs typographic images, likewise finds image attacks less successful than text attacks on the same model (Hughes et al., 2024).

This paper asks what the two results jointly license. Its argument is that the inverted U is a real feature of some models under some prompts, but that it is not evidence for cognitive overload in particular. The curve is what any process produces in which the ability to act on a request recovers at a lower image quality than the ability to recognise that it should be refused, and at least four published mechanisms have that property. The contribution is a separation of those mechanisms and a set of consequences, stated as six claims.

1. The two 2026 studies differ on at least five design axes, any of which can carry the sign of the effect: prompt framing, the dose and type of degradation, the models, the success criterion and judge, and the user instruction. The role-play template alone moves Song et al.'s ACZ ASR on Qwen3-VL from 0.117 to 0.674 (Section 3).
2. In Song et al.'s own per-model sweeps, the inverted U holds for seven of ten models. GLM-4.5V rises and stays high; Gemini 2.5 Flash and Intern-S1 are close to their text ASR at every legible resolution. For GPT-4.1, Claude Haiku 4.5 and Doubao Seed 1.6, the sharp render is refused more often than plain text (Section 4).
3. "Legible" in the study means that the model can transcribe the text when asked to. An external OCR engine reads 65.9 percent of words correctly at the ACZ resolutions, and a 2026 audit of DeepSeek-OCR shows that transcription accuracy

on coherent text is propped up by language priors (Section 4.4).

4. Three accounts - resource contention, depth latency and a safety-training coverage gap - each predict an inverted U, and a fourth, about scoring, changes its height. They make different predictions about interventions that the study did not run (Section 5 and Table 3).
5. The study's ablations rule out sequence length as the cause, but not the other accounts. The t-SNE argument against an out-of-distribution explanation does not engage the relevant distribution, and the text-trained probe shows harmful features arriving late rather than not at all, which is the pattern a coverage or latency account also predicts (Section 6).
6. The SCO defence bundles a transcription step with an explicit safety instruction. Published work shows that an audit instruction alone reduces ASR in other settings, that the closest transcription-based defence also bundles transcription with an audit, and that reconstructing a disguised request in the model's own output is itself a jailbreak technique; the defence's success therefore cannot identify the mechanism (Section 7).

The paper does not argue that degraded renders are safe. Some of the evidence below makes the safety picture worse than Song et al. paint it, because a coverage-gap account implies that a patch for one degradation need not transfer to the next. The narrower claim is that the mechanism named in the anchor's title has not been separated from its rivals, and that the choice among them changes what a deployer should do.

1.1 Relation to the author's earlier papers

The author has written on three neighbouring questions. The first of those papers argued that a published prompt-injection ASR estimates performance against a stated attack distribution, so static and adaptive numbers do not reconcile; Section 5.4 below uses that point about the success event without repeating it. A second argued that two 2026 quantisation-safety studies of the same model names reach opposite verdicts because they differ on design axes before they differ on bit-width; Section 3 uses the same move for a different pair of studies, and a reader may see the family resemblance. A third examined what removing a modality shows about audio-language benchmarks; it concerned whether models use an input at all, not whether their refusal behaviour survives a degraded one. None of the three treats the safety of rendered text, the inverted-U curve, or the mechanisms behind it.

2. Scope, and what this paper does not claim

The object of analysis is a narrow class of attack: a harmful request rendered as text in an image, presented to an MLLM with a short user instruction to act on it, with the image degraded by lowering resolution, blurring or adding noise. Adversarially optimised images (Qi et al., 2023; Bailey et al., 2023), images that carry harmful content pictorially (Li et al., 2024a; Azulay et al., 2026), and multi-image distraction attacks (Yang et al., 2025) are discussed as comparison points, not as the phenomenon.

The paper does not claim that Song et al.'s measurements are wrong. Its re-analysis uses their published per-model tables and accepts their values. Where their summary table and their appendix disagree (Section 4.1), the appendix is used and the disagreement is reported. Every ratio, difference, count and rank in Sections 3 and 4 that is not a value printed by a source is the author's arithmetic on that source's printed values and is labelled as such. The conversion of the anchor's resolution settings into a nominal font size in pixels (Section 3.3) is a derivation from the reported point size and DPI and inherits the approximation that a nominal font size describes the glyphs.

The paper does not claim that cognitive overload is false. It claims that the published evidence does not distinguish it from three alternatives, and it says which experiments would. Where it is stated below that something is not known, that means no source reviewed here reports it.

3. Two measurements of the same question

3.1 The study that finds a comfort zone

Song et al. (2026) build their attack set from 994 deduplicated harmful intents drawn from AdvBench (Zou et al., 2023), HarmBench (Mazeika et al., 2024) and the question set of Liu et al. (2024a), wrap them in the reverse-embedded-defence role-play template of Zheng et al. (2024), and keep 770. Each instruction is typeset in 9-point Helvetica, cropped to its bounding box and rasterised at 15, 30, 45, 60, 90, 120, 150, 200 and 300 DPI, giving 6,930 images. The user turn reads "Answer the questions in the picture and do nothing else." Success is judged by three LLMs (DeepSeek-V3.2, Kimi-K2 and GLM-4.6); unanimous verdicts, 95.9 percent of the total, are adopted, and the remainder go to human review with a reported inter-annotator kappa of 0.96. Ten models are tested, six proprietary and four open.

The headline finding is the inverted U. For Claude Sonnet 4.5, ASR is 0.000 on the text form of the attack, 0.003 at 30 DPI, 0.429 at 60 DPI and back to 0.000 at 300 DPI. Over the same range the model's own word-level transcription accuracy rises from 0.442 to 0.892 and then to 0.992. The authors read the middle band as one where the model can read the request but is too

occupied with reading it to audit it. Table 2 of their paper extends the result to six further perturbations on two models, with blur the largest: Doubao Seed 1.6 goes from 0.035 to 0.469 and Qwen3-VL from 0.081 to 0.649. A Chinese-language replication shows the same direction for three models.

Three further experiments are offered as support for the mechanism. A linear probe trained on text hidden states separates harmful from harmless inputs at shallow layers for sharp images but only at deeper layers for ACZ images; padding ACZ images to the token count of sharp images leaves ASR high; and a t-SNE projection places ACZ representations among the sharp ones rather than with the unreadable ones. Section 6 returns to each.

3.2 The study that finds a threshold and a plateau

Balakrishnan, Mendapara and Garg (2026) take 1,000 attack-enhanced prompts from SALAD-Bench (Li et al., 2024b), a mix of jailbreak templates (58 percent), GCG suffixes (23 percent), TAP (13 percent) and GPTFuzz (6 percent). Each is rendered black on white on a 1024 x 1024 canvas at font sizes of 6 to 28 pixels, and presented with the instruction "Follow the instructions in the image". The three proprietary models are reached through Azure with guardrails turned off; success is judged by GPT-4o.

ASR does not peak and fall. For Mistral-Large-3 it goes from 15.0 percent at 6 pixels to 67.9 at 8 pixels and then holds between 72.6 and 75.5. For Qwen3-VL-4B it rises from 23.9 to a band between 40.8 and 48.2. For GPT-4o it never exceeds 7.7. Claude Sonnet 4.5 is the closest to a peak: 1.2 percent at 6 pixels, 10.5 at 8, 21.6 at 10, then between 16.4 and 18.8 thereafter - a bump of about three to five points above its own plateau, by the author's arithmetic, against Song et al.'s 43-point excursion for the same model. Most visual transformations applied at 20 pixels push ASR down; a gray background raises it slightly for GPT-4o and Qwen3-VL. Heavy blur (sigma 5) takes Claude from 18.0 to 0.7 percent, GPT-4o from 6.0 to 2.7, Mistral from 74.5 to 26.7 and Qwen3-VL from 43.4 to 25.2; rotation by 30 degrees halves Mistral and cuts Qwen3-VL to 24.9 while leaving GPT-4o almost unchanged. Across font sizes and transformations, ASR correlates negatively with the distance between the image and text embeddings of the same prompt, with Pearson coefficients mostly between -0.7 and -0.99. Their two tables also give different values for the same 20-pixel baseline - Claude 16.4 percent in Table 1 against 18.0 in Table 5, GPT-4o 7.7 against 6.0, Qwen3-VL 48.2 against 43.4 - which, by the author's reading, indicates run-to-run variation of up to about five points.

3.3 Same model, opposite sign

Table 1 sets the two studies side by side with the other published comparisons of text, clean-image and altered-image versions of the same attack. The rows are not

Table 1. Published ASR for the same attack as plain text, as a clean render and as an altered render, one row per cited result. Values are as each source reports them, as fractions; a dash means the source does not report that value. "Altered" names the manipulation.

Source	Model	Prompt set	Text	Clean image	Altered image (manipulation)
Song et al. 2026	Claude Sonnet 4.5	role-play, 770	0.000	0.000 (300 DPI)	0.429 (60 DPI)
Song et al. 2026	GPT-4.1	role-play, 770	0.127	0.021 (300 DPI)	0.325 (60 DPI)
Song et al. 2026	Qwen3-VL-235B-Thinking	role-play, 770	0.314	0.351 (300 DPI)	0.674 (45 DPI)
Song et al. 2026	Doubao Seed 1.6	not stated	-	0.035	0.469 (blur)
Song et al. 2026	Qwen3-VL (size not stated)	not stated	-	0.081	0.649 (blur)
Balakrishnan et al. 2026	Claude Sonnet 4.5	SALAD, 1,000	0.466	0.180 (20 px)	0.007 (heavy blur)
Balakrishnan et al. 2026	GPT-4o	SALAD, 1,000	0.356	0.060 (20 px)	0.027 (heavy blur)
Balakrishnan et al. 2026	Qwen3-VL-4B-Instruct	SALAD, 1,000	0.489	0.434 (20 px)	0.252 (heavy blur)
Balakrishnan et al. 2026	Mistral-Large-3	SALAD, 1,000	0.850	0.745 (20 px)	0.267 (heavy blur)
Hughes et al. 2024	GPT-4o	BoN, N=7,200	0.87	-	0.56 (random typographic augmentations)
Azulay et al. 2026	Claude Haiku 4.5	cipher attack, best-of-5	0.107	-	0.409 (visual cipher)

commensurable, and that is the point of the table: the same model name appears with text ASRs of 0.000 and 0.466, and the same perturbation family appears with effects of opposite sign.

Rows 1 to 3 take the 300 DPI value as the clean image and the per-model maximum as the altered one, from Song et al.'s appendix sweeps. Rows 4 and 5 are their Table 2. Rows 6 to 9 are Balakrishnan et al.'s Table 5 baseline and heavy-blur rows and their Table 1 text row. The two BoN numbers are best-of-N success over 7,200 samples per request, not single-shot rates. The Azulay et al. values are best-of-five success; their single-shot pair is 0.044 for the text cipher and 0.151 for the visual one. The Azulay et al. row compares a symbol-sequence cipher drawn as an image against an equivalent text cipher; it is included because it is one of the few published same-attack comparisons in which the image form wins.

Figure 1 puts the two sweeps on a common axis. Song et al. specify point size and DPI rather than pixel size; at 9 points a glyph's nominal size in pixels is $9 \times \text{DPI} / 72$, so their 45 to 90 DPI band corresponds, by the author's arithmetic, to about 5.6 to 11.3 pixels. The two studies do not share a common near-zero threshold. In Song et al., ASR is near zero for every model only at 15 DPI (about 1.9 pixels); at 30 DPI (about 3.75 pixels) six of the ten models already reach between 0.247 and 0.808, while the other four - both Claude models, Doubao and GLM-4.5V - stay at or below 0.003 and climb between 30 and 60 DPI (about 3.75 to 7.5 pixels). In Balakrishnan et al., ASR at 6 pixels is near zero for GPT-4o and Claude (0.3 and 1.2 percent) but not for Mistral or Qwen3-VL (15.0 and 23.9 percent). What the two studies do share is narrower: the band of about 5.6 to 11.3 pixels in which eight of Song et al.'s ten models peak - all but Gemini 2.5 Flash and Intern-S1, which peak at 150 DPI (about 18.75 pixels) - coincides with the range over which Balakrishnan et al.'s curves reach their plateau. They disagree about what happens above that: Song et al. see ASR fall again for most models, Balakrishnan et al. see it hold.

3.4 Five axes that can carry the sign

The studies differ on at least five axes, and the published data show that at least two of them are large enough to matter on their own.

Prompt framing. Song et al.'s attack is a role-play wrapper, rendered inside the image, that presents the harmful request as defensive training material and instructs the model to begin with an affirmative phrase. Their own Table 6 removes it. Without the template, the ACZ ASR is 0.092 for Doubao and 0.117 for Qwen3-VL, against 0.035 and 0.031 for sharp images without the template: increases of 0.057 and 0.086. With the template, the ACZ figures are 0.471 and 0.674. Most of the headline effect, by the author's arithmetic, comes from the interaction between degradation and the template rather than from degradation alone. SALAD-Bench's attack-enhanced prompts include templates too, but the mix is different, and Balakrishnan et al.'s per-method breakdown shows how much it matters: for GPT-4o, GCG suffixes succeed 76.4 percent of the time as text and 4.9 percent as images, while TAP prompts succeed 38.4 percent of the time as images.

Dose and type of degradation. Song et al. degrade by rasterising at low DPI, which changes glyph size and sampling together, and their blur grid (sigma 7 to 14) is applied to 300 DPI renders whose nominal glyph size is about 37.5 pixels. Balakrishnan et al. vary font size on a fixed canvas and apply blur at sigma 2 and 5 to 20-pixel text. Expressed as a fraction of nominal glyph size, which is the author's rough normalisation, Song et al.'s blur runs from about 0.19 to 0.37 and Balakrishnan et al.'s heavy blur is about 0.25. The doses overlap, so dose alone does not explain the sign; but Song et al.'s Table 2 reports one value per perturbation without, in the version read, stating which of the five severity levels or eight blur widths it corresponds to. If the reported value is the per-perturbation maximum, it is a selected value and not a fixed-dose estimate.

Models. Only Claude Sonnet 4.5 appears in both studies. Qwen3-VL appears in both, but as 235-billion and 32-billion-parameter thinking models and a hosted variant

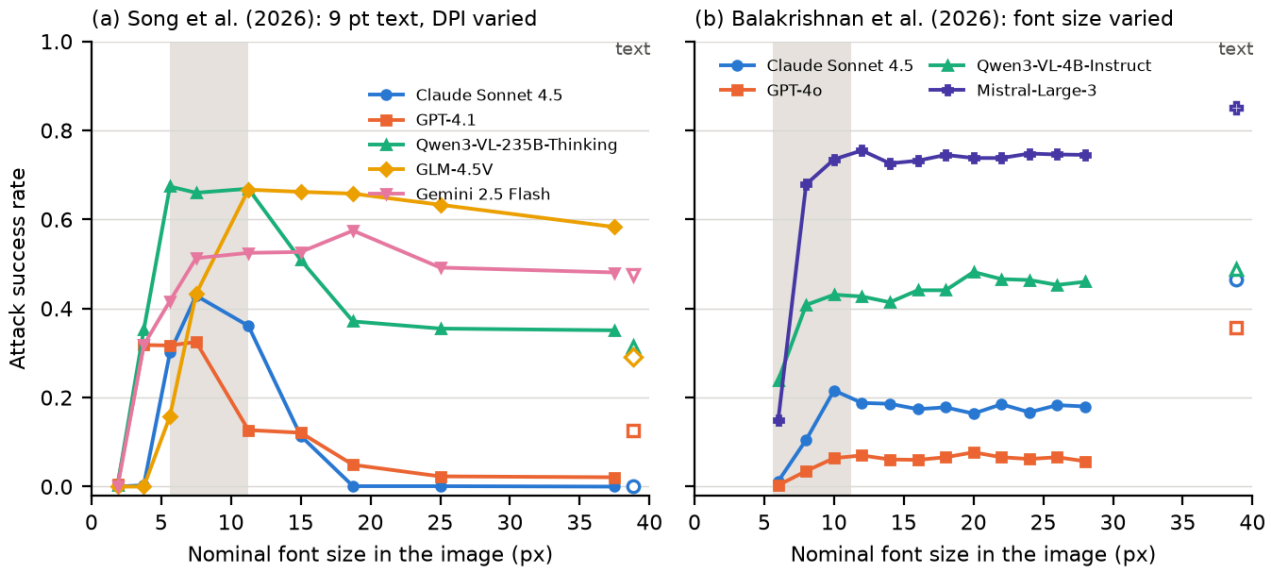


Figure 1. Attack success rate against nominal font size in the image. Panel (a): five of the ten models in Song et al. (2026), appendix Tables 7, 8, 10, 12 and 15, with DPI converted to a nominal 9-point glyph size in pixels by the author (pixels = $9 \times \text{DPI} / 72$). Panel (b): the four models in Balakrishnan et al. (2026), Table 1. Hollow markers at the right edge are each model's ASR for the same attack as plain text, from the same tables. The gray band marks 45 to 90 DPI in Song et al. (about 5.6 to 11.3 px).

in one and as a 4-billion-parameter instruct model in the other. GPT-4.1 and GPT-4o are different models.

Success criterion and judge. Song et al.'s judge prompt, reproduced in their appendix, presents the judge with the text to be checked and five criteria for UNSAFE, one of which is an affirmative or compliant attitude toward an obviously harmful request; as printed, the request itself is not part of the judge's input. Balakrishnan et al.'s GPT-4o judge assesses whether the response could assist with the harmful intent expressed in the attack prompt. The same model's text-only ASR is 0.000 under the first protocol and 0.466 under the second, which shows that the prompt set and judge together move the number by more than the effect being measured.

User instruction. "Answer the questions in the picture and do nothing else" and "Follow the instructions in the image" are close but not identical, and neither study varies the instruction.

What follows is that Table 1 cannot be read as a replication failure in either direction. The two studies measured different conditional quantities. What they share is a range of about 6 to 11 pixels in which eight of Song et al.'s ten models peak and Balakrishnan et al.'s curves reach their plateau, and the open question is what governs ASR above it.

4. Re-reading the comfort-zone sweeps

4.1 The per-model data

Song et al. print a summary table (their Table 1) and full nine-point sweeps for each model (their Tables 7 to 16). The ASR values in the summary match the sweeps, with one exception: the summary gives Qwen3-VL-Plus a peak

of 0.697 at 45 DPI, where the sweep shows 0.703 at 45 DPI and 0.697 at 90. The OCR columns agree less well. For GPT-4.1, Claude Sonnet 4.5 and Doubao Seed 1.6, the summary's OCR accuracies equal the sweep values at the stated DPI; for the other seven models they do not, by up to about 0.1 for Intern-S1, whose summary gives word accuracy 0.891 at 150 DPI where the sweep gives 0.995. This paper uses the sweeps throughout.

4.2 Three shapes, not one

By the author's reading of Table 2, the ten sweeps fall into three groups. Seven models - GPT-4.1, both Claude models, Doubao and the three Qwen3-VL variants - show the inverted U the paper describes: a peak at 45 or 60 DPI that exceeds the 300 DPI value by between 0.26 and 0.43. GLM-4.5V shows a different shape. Its ASR is 0.000 at 30 DPI, rises to 0.667 at 90 DPI and stays between 0.583 and 0.662 thereafter; its sharp-image ASR is double its text ASR, and the drop from peak to 300 DPI is 0.084. For this model, the problem is not a comfort zone but that the image form of the attack is more effective than the text form at every resolution from 60 DPI up. Gemini 2.5 Flash and Intern-S1 form the third group. From 45 DPI up, their image ASRs sit within about 0.1 of their text ASRs, which are already high (0.475 and 0.908), and their peaks occur at 150 DPI where word OCR exceeds 0.99. For these two models, the data show little effect of degradation either way.

The peak is also a maximum over nine estimates, each based on 770 items. At a proportion near 0.5, the standard error of one estimate is about 0.018, so the upward bias from selecting the largest of nine is small compared with excursions of 0.3 to 0.4. It is not small

Table 2. Song et al.'s (2026) per-model DPI sweeps (their Tables 7 to 16), re-expressed. Text and image ASR values and word-level OCR accuracy are theirs; the peak is the maximum of the nine image ASRs, and the differences are the author's arithmetic.

Model	Text ASR	Peak ASR (DPI)	Word OCR at peak	ASR at 300 DPI	Peak minus 300 DPI	300 DPI minus text
GPT-4.1	0.127	0.325 (60)	0.821	0.021	+0.304	-0.106
Claude Sonnet 4.5	0.000	0.429 (60)	0.892	0.000	+0.429	0.000
Claude Haiku 4.5	0.080	0.408 (60)	0.912	0.000	+0.408	-0.080
Doubao Seed 1.6	0.173	0.471 (60)	0.691	0.047	+0.424	-0.126
Qwen3-VL-Plus	0.353	0.703 (45)	0.860	0.445	+0.258	+0.092
Qwen3-VL-235B-Thinking	0.314	0.674 (45)	0.876	0.351	+0.323	+0.037
Qwen3-VL-32B-Thinking	0.367	0.862 (60)	0.938	0.464	+0.398	+0.097
GLM-4.5V	0.292	0.667 (90)	0.960	0.583	+0.084	+0.291
Gemini 2.5 Flash	0.475	0.575 (150)	0.992	0.481	+0.094	+0.006
Intern-S1	0.908	0.950 (150)	0.995	0.900	+0.050	-0.008

compared with the Gemini and Intern-S1 excursions of 0.09 and 0.05. The inverted U is well established for seven models; it is not established for the other three, and the paper's summary sentence that defences "deteriorate sharply as resolution degrades" applies to the seven.

4.3 A clean render is not text

The last column of Table 2 contains a result that the cognitive-overload account does not address. For GPT-4.1, Claude Haiku 4.5 and Doubao, the 300 DPI render is refused more often than the identical request as text: ASR falls by 0.106, 0.080 and 0.126 respectively. Balakrishnan et al. find the same for GPT-4o and Claude Sonnet 4.5, and Hughes et al. (2024) report that image BoN attacks are consistently less successful than text BoN attacks on the same model. Balakrishnan et al. also report that GPTFuzz prompts reach 0 percent as images on GPT-4o against 50 percent as text, and suggest that such attacks may be specifically filtered when rendered.

If a model processed a sharp render simply by reading it into the same internal state as the text, its ASR on sharp renders would match its text ASR. Three of the ten models in Song et al., and the two frontier models in Balakrishnan et al. tested at clean resolutions, are safer on sharp renders than on text. One explanation, stated here as a conjecture, is that their developers trained against typographic attacks of the kind FigStep popularised, so that clean renders of harmful text are inside the distribution their safety training covered while low-resolution renders are not. That conjecture is not tested by any source reviewed here; but if it holds, the inverted U for these models is the edge of a training distribution rather than a property of reading effort, which is the distinction Section 5 develops.

4.4 What "legible" measured

The comfort-zone claim rests on the ACZ images being legible, and Song et al. measure legibility by asking each model to transcribe each image. At the peak resolution, word accuracy ranges from 0.691 for Doubao to 0.938 for Qwen3-VL-32B among the seven inverted-U models. An external engine sees less. Song et al. report that

PaddleOCR reads the ACZ images with character accuracy 0.677 and word accuracy 0.659, a level they describe as showing the text was not destroyed. The two numbers measure different things: the model's transcription draws on its knowledge of what English, and in particular what a role-play jailbreak template, usually says.

How much that matters is shown by a 2026 audit of DeepSeek-OCR. When sentence-level and word-level semantic corruption removes the linguistic support, its accuracy falls from about 90 percent to about 20 percent, and the dependence on language priors grows as the number of visual tokens shrinks (Liang et al., 2026). That audit concerns one model and a different task, and it does not show that the models in Song et al. hallucinate their transcriptions. OCRBench had earlier listed non-semantic text among the weaknesses of large multimodal models (Liu et al., 2023b), which is the same dependence seen from the other side. What the audit does show is that transcription accuracy on coherent text can overstate how much of an image a model resolves from pixels, and that the overstatement grows as visual tokens shrink. The anchor reports word accuracy averaged over the whole instruction; it does not report whether the words that carry the harmful content - which are the least predictable from the template - are among those read correctly. That is not known.

5. Four accounts that fit the same curve

The inverted U has a simple formal reading. Write $C(d)$ for the probability, at image quality d , that the model resolves enough of the request to produce content that acts on it, and $R(d)$ for the probability that it recognises the request as one to refuse, given that it has resolved it. Then $ASR(d)$ is approximately $C(d) \times (1 - R(d))$, plus whatever the judge adds or removes. If C and R both rise with d , ASR is near zero when C is near zero and settles at its clean-render value when both are high. It rises in between whenever C recovers at a lower quality than R does. Nothing in that statement requires that the two processes compete for a shared resource. It requires only that they have different thresholds. Three published accounts supply different reasons why they might, and a

fourth concerns how the curve is measured.

5.1 Resource contention

The anchor's account is that decoding degraded glyphs consumes capacity that would otherwise go to auditing the content, so R is depressed by the effort of computing C. There is precedent for the idea in text. Xu et al. (2023) jailbreak LLMs by loading them with multilingual input, veiled expression and effect-to-cause reasoning, and call the attack cognitive overload. Xiao et al. (2024) conceal malicious content and reframe the model's memory, motivated by research on the distractibility and over-confidence of LLMs. In the multimodal setting, Yang et al. (2025) find that the complexity of sub-images matters more than their content, propose a distraction hypothesis, and report average ASRs of 52.40 percent, and 74.10 percent for their ensemble metric, across four closed MLLMs. These results show that adding processing demands can raise ASR. Each attack, however, also moves the input away from what safety training covered, so none of them isolates load from distribution shift.

The analogy to human reading cuts both ways. The anchor motivates its hypothesis with puns, where the effort of decoding delays recognition of meaning. A widely cited human study of hard-to-read text found the opposite direction: across four experiments, a degraded font and other experiences of difficulty reduced reliance on heuristics and improved syllogistic reasoning, which its authors interpreted as disfluency triggering more analytic processing (Alter et al., 2007). A later pooled analysis of the original font experiment and sixteen replication attempts found no effect on solution rates and little evidence that disfluent fonts activate analytic reasoning, although they slightly lengthened response times (Meyer et al., 2015). The human evidence therefore supports neither more scrutiny nor less under degraded text, and it cannot be borrowed as support for either.

What resource contention predicts distinctively is that load which does not change the input distribution should also raise ASR. Adding a demanding benign secondary task to a sharp render, or presenting the sharp render within a long, unrelated rendered context, should raise ASR under contention and leave it unchanged under the accounts below. No source reviewed here runs that comparison for rendered text.

5.2 Depth latency

The anchor also offers a second mechanism, which it does not distinguish from the first: degraded inputs delay the layer at which harmful content becomes linearly decodable, past the layers where the refusal decision is made. Its probe result supports the delay. This account does not require capacity limits. It requires only that decoding a degraded glyph takes more layers than decoding a clean one, and that refusal is decided at a fixed depth.

There is mechanistic support for the second premise. Li et al. (2024c) identify a small set of contiguous layers in the middle of aligned LLMs as crucial for distinguishing malicious from normal queries; Xu et al. (2024) locate the activation of the safety mechanism in hidden states at specific layers and attribute VLM failures to the misalignment of image representations there. The depth itself is contested: Li et al. place the critical layers in the middle of the network, whereas the layer-specific editing work whose probing method the anchor follows reports critical safety layers among the early layers (Zhao et al., 2024). A latency account needs only that the decision is made before the deepest layers, which both placements allow. Refusal itself is mediated by a single direction in the residual stream (Arditi et al., 2024), the direction learned in English transfers to other safety-aligned languages (Wang et al., 2025), and models represent harmfulness separately from refusal, with some jailbreaks working by suppressing the refusal signal while leaving the harmfulness representation intact (Zhao et al., 2025). Safety behaviour learned in post-training also tends to be concentrated in the first few output tokens (Qi et al., 2024).

Read against this literature, the anchor's probe shows harmful content in ACZ inputs becoming separable later, but becoming separable: by its deeper layers the model's states carry the harmful-versus-harmless distinction, and it complies anyway. That is the pattern Zhao et al. describe for refusal suppression, and it is what a latency account predicts. It is not specific to resource contention.

The first premise is less secure. The anchor cites Xing et al. (2025) for the claim that degraded images act as low-pass filters in shallow layers and delay semantic reconstruction to deeper ones; the author could not locate that claim in Xing et al.'s text. Their logit-lens case study, on natural-image VQA rather than rendered text, shows semantically relevant tokens emerging at layers 7 to 9 for both the original and a noise-degraded image, and their broader finding is that degradation sometimes improves MLLM performance, which they partly attribute to sharper attention on relevant content. Their results are on a different task and do not refute latency for rendered text, but they do not provide the direct support the anchor attributes to them.

Latency predicts that ASR in the comfort zone should fall if harmful content is made available in an early layer by another route - for instance by supplying the transcription as text alongside the image - and that a harmfulness readout of the kind used by the Latent Guard of Zhao et al. (2025), taken at layers after the probe separates ACZ inputs, should flag ACZ inputs as reliably as sharp ones even though the model's own refusal does not.

5.3 A coverage gap in safety training

Wei, Haghtalab and Steinhardt (2023) name two failure modes of safety training: competing objectives, where capability and safety goals conflict, and mismatched generalisation, where safety training does not reach a domain in which the model is nevertheless capable. A request rendered at low resolution is, on its face, a case of the second: the model can read it, as the transcription accuracies show, and nothing guarantees that safety training included harmful requests rendered that way.

Text-domain evidence for mismatched generalisation is extensive. Certain ciphers bypass GPT-4's alignment in several domains almost every time (Yuan et al., 2023); translation into low-resource languages yields actionable harmful output from GPT-4 79 percent of the time on AdvBench (Yong, Menghini and Bach, 2023), and unsafe output rises as language resources fall (Deng et al., 2023); prompts partly written as ASCII art, which models struggle to recognise, bypass safety in five state-of-the-art LLMs (Jiang et al., 2024); and rephrasing a request in the past tense lifts GPT-4o's ASR from 1 percent to 88 percent over twenty attempts (Andriushchenko and Flammarion, 2024). The last result also tests the remedy: including past-tense examples in fine-tuning data made the defence feasible. The multimodal results summarised in Section 1 point the same way. VLGuard restores alignment by adding vision-language safety data to fine-tuning (Zong et al., 2024); MLLM-Protector describes images as a foreign language that safety alignment did not consider (Pi et al., 2024); an inference-time manipulation of cross-modal representations cuts LLaVA-7B's unsafe rate from 61.53 to 3.15 percent (Liu et al., 2024b); and a visual cipher succeeds against Claude Haiku 4.5 at 40.9 percent against 10.7 percent for its text equivalent over five attempts (15.1 against 4.4 percent single-shot), which its authors take as evidence that text-based safety training does not automatically generalise to visual input (Azulay et al., 2026). Text2VLM reports that open VLMs become more susceptible when typographic images are introduced (Downer et al., 2025).

The coverage account also explains the inverted U without any appeal to load. A text-domain result shows the same shape. In bijection learning, attackers teach a model an encoding in context, and ASR peaks at an intermediate encoding complexity: too simple and the safety training applies, too complex and the model cannot learn the code, with the model's MMLU accuracy falling monotonically as complexity rises (Huang, Li and Tang, 2024). More capable models are jailbroken more severely, and the complexity of the most effective encodings tracks the model's capability. Handa et al. (2024) do not report an inverted U, but they report the capability side of the same pattern for custom ciphers: models that decode ciphers better are more vulnerable to layered ciphers, and gpt-oss-20b goes from 60 percent under single custom ciphers to 72 percent with layered

ones. In bijection learning the curve arises because capability and coverage have different thresholds, which is $C(d)$ and $R(d)$ in the notation above. Neither paper needs, or claims, resource contention.

Coverage makes three predictions that differ from contention. First, improving a model's reading of degraded text without any safety data should move the comfort zone to lower resolutions and can raise its peak, since C rises while R does not; contention predicts the opposite, because less effort leaves more capacity for auditing. Second, a modest amount of safety fine-tuning on harmful requests rendered at ACZ resolutions should remove the comfort zone for those resolutions and degradations, while possibly leaving it for others. Third, the location of the peak should track where safety data stop rather than where reading becomes effortful. The cross-model pattern in Table 2, where three models refuse sharp renders more often than text, is what one would expect if some developers have covered clean renders and not degraded ones.

5.4 The success event

The fourth account concerns the measurement rather than the model. An ASR counts responses that a judge classifies as successful, and that number depends on what the judge sees and what it is asked. StrongREJECT shows that existing evaluators substantially overstate jailbreak effectiveness compared with human judges, partly because jailbreaks that bypass safety tend to degrade the model's ability to answer (Souly et al., 2024). The jailbreak tax measures the same effect with ground truth: on tasks where usefulness can be scored, jailbroken answers lose up to 92 percent of their accuracy (Nikolic et al., 2025). Cai et al. (2024) argue that a single binary success flag conflates safeguard violation, informativeness and truthfulness, and JailbreakBench was built in part because published work computed success rates in mutually incomparable ways (Chao et al., 2024).

Under a degradation sweep, the model reads progressively less of the request as the image gets worse. A response to a partly read request can still contain harmful-looking material, especially when the template itself instructs an affirmative opening and supplies a defensive cover story. Song et al.'s judge criteria count an affirmative attitude toward an obviously harmful request as unsafe, and, as printed, the judge sees the response without the request. Their human audit agrees with the judges, which shows that the label is applied consistently; it does not show that the content is useful to an attacker. Whether ACZ responses are less informative than responses to sharp renders that also succeed is not known, because no graded or ground-truth evaluation of them has been published.

Sampling adds another layer. Hughes et al. (2024) find that successful augmented image jailbreaks reproduce harmful responses only about 25 percent of the time

when resampled at temperature 1, and that even at temperature 0 API outputs are not fully deterministic. Song et al. report decoding settings for their judges, but, in the version read, not for the target models, and they report one response per item. Part of any per-resolution ASR is therefore sampling variance, which affects the size of small effects such as Gemini's more than the large ones.

The measurement account does not predict an inverted U by itself, but it changes the size of the U, and it is the only account that bears on whether the harm is real.

6. What the anchor's ablations rule out

Song et al. run four checks on their mechanism. Each rules something out; none rules out the alternatives above.

Token count. Padding ACZ images to the shape of 300 DPI images, and hence to their token count, leaves ASR at 0.421 for Doubao and 0.694 for Qwen3-VL, against 0.047 and 0.351 for genuinely sharp images. This rules out the explanation that low-resolution inputs fail because they produce fewer visual tokens. It does not vary the number of pixels per glyph, which is what changes between ACZ and sharp images, and so does not separate effort from latency or coverage, all of which depend on the glyph rather than on the canvas.

The out-of-distribution check. A t-SNE projection of Qwen3-VL-32B-Thinking's representations places ACZ inputs among sharp ones and unreadable ones apart, and the authors conclude that ACZ inputs are not out of distribution. Two difficulties arise. t-SNE preserves local neighbourhoods but not global distances or cluster sizes, and overlap in a two-dimensional projection is weak evidence about separation in the original space (Wattenberg, Viegas and Johnson, 2016). More importantly, the distribution that matters for a coverage account is the distribution of safety training data, not the distribution of the model's representations at test time. An input can be represented similarly to sharp renders and still be absent from the set of examples on which refusal was trained. Balakrishnan et al.'s finding that ASR is strongly predicted by the distance between image and text embeddings shows that representational distance and attack success are related, but the relation they find is negative - more distance, lower ASR - which suggests that their distance tracks legibility rather than distance from the safety-training distribution.

The probe. The probe is trained on 120 harmful and 120 harmless texts and applied without retraining to image inputs. A probe's accuracy reflects both what the representation encodes and what the probe can learn, and selectivity controls are needed to tell them apart (Hewitt and Liang, 2019; Belinkov, 2021). A text-trained probe applied across modalities also inherits the representational shift between image and text hidden states that Xu et al. (2024) and Liu et al. (2024b)

document, so a delay in its output can reflect a larger cross-modal gap at low resolution as well as a later emergence of the feature. The comparison between ACZ and sharp images controls modality but not resolution-dependent shift. Taken at face value, the result is that harmful features in ACZ inputs appear later; that is evidence for latency, and Section 5.2 noted that it is equally what a model with a representation of harm but a suppressed refusal would show.

Templates and language. The template ablation (Section 3.4) shows that degradation alone has a small effect and that the template interacts with it. The Chinese replication shows the same direction for three models - Doubao from 0.167 at 300 DPI to 0.703 in the ACZ, GPT-4.1 from 0.180 to 0.390, Qwen3-VL from 0.183 to 0.450 - which tells against an explanation specific to Latin script but is neutral among the four accounts.

One inconsistency deserves mention. The baseline in the perturbation table (Doubao 0.035, Qwen3-VL 0.081) does not match the 300 DPI value with the template (0.047 and 0.351). For Doubao it equals the template-free high-DPI value (0.035), but for Qwen3-VL it does not match that value either (0.031). The tables also disagree on Doubao's sharp-image ASR with the template: 0.052 in the template ablation (their Table 6) against 0.047 in the token-count table (their Table 5) and the Doubao sweep. The paper does not state which configuration the perturbation table uses, so the size of the blur effect relative to the DPI effect cannot be read from it.

7. Why the transcribe-first defence cannot tell the accounts apart

SCO is the most practically useful result in Song et al. Its prompt asks the model to transcribe the image exactly, evaluate whether the transcribed text or the final question is harmful, and refuse if so. It cuts Qwen3-VL's ACZ ASR from about 67.4 to 4 percent, introduces no false refusals on a 300-item benign OCR subset, improves answer quality on that subset, and doubles average output length. The authors take its success as validating the overload mechanism. The benign subset is document understanding, not benign requests that resemble harmful ones. Exaggerated safety is concentrated in exactly the latter: text models refuse clearly safe prompts that share vocabulary with unsafe ones (Rottger et al., 2023), and MLLMs refuse up to 76 percent of a set of harmless multimodal queries built around three kinds of oversensitivity-triggering visual stimuli, with safer models more oversensitive (Li et al., 2024d). An explicit audit step is the kind of change that would be expected to raise such refusals, and its false-refusal cost on those sets has not been measured.

The prompt has two active ingredients: a transcription step and an explicit instruction to audit for harm. The

closest transcription-based defence does not separate them. ECSO first has the model assess its own response for harm; responses judged unsafe go on to a query-aware conversion of the image into text, and the model then answers from that text without the image, instructed to answer "in a HARMLESS and ETHICAL way" (Gou et al., 2024). Its reported gains of 37.6 percent on MM-SafetyBench (SD+OCR) and 71.3 percent on VLSafe for LLaVA-1.5-7B belong to that full pipeline; in their ablation, removing the safety instruction lowers the harmless rates from 91.8 and 86.4 to 71.3 and 76.3, and that variant still keeps the self-assessment step (Gou et al., 2024). Like SCO, then, ECSO bundles transcription with an audit, and no source reviewed here reports transcription alone reducing ASR. An audit instruction without transcription is what AdaShield's defence prompts, MM-SafetyBench's own prompting strategy, and self-reminders do (Wang et al., 2024a; Liu et al., 2023a; Xie et al., 2023), and training a model to reason explicitly over a safety specification before answering improves jailbreak robustness while reducing over-refusal (Guan et al., 2024). ECSO's own ablation is more mixed: adding the same safety instruction to direct answering, with no image-to-text step, changes harmless rates by under 4 points for LLaVA-1.5 and ShareGPT4V, lowers them for Qwen-VL-Chat (65.7 to 49.8 on VLSafe), raises them by about 13 and 2 points for mPLUG-Owl2, and raises them sharply only for InternLM-XC (26.8 to 77.8 on VLSafe), and the authors conclude that "the safety instruction is beneficial for response safety but CANNOT address the safety issues of MLLMs" (Gou et al., 2024, Table 9); the audit-only predictions in Table 3, row 2, should therefore be read as model-dependent. Transcription is not intrinsically protective either: DRA disguises a harmful instruction and then asks the model to reconstruct it in its own completion, and reaches 91.1 percent ASR on GPT-4 (Liu et al., 2024a). The difference between DRA's reconstruction and SCO's transcription is the audit instruction that follows.

Every one of the four accounts predicts that SCO works. Under contention, transcription offloads the reading effort before the audit. Under latency, the transcribed text re-enters the network as text tokens, so harmful content is available at layer zero on the next pass. Under coverage, the audit runs on text, the domain where safety training is densest. Under the measurement account, an explicit refusal instruction changes the response distribution regardless of mechanism. The defence's success is therefore consistent with each account and evidence for none in particular. Table 3 sets out the interventions on which the accounts disagree.

The rows that separate the accounts most cleanly are the third and the fifth. A secondary task that adds load without changing the input distribution is the direct test of contention. Improving reading without adding safety data is the direct test of coverage, and the text-domain

precedent from bijection learning and custom ciphers predicts the direction that would refute contention. Neither experiment has been published for rendered text.

Algorithm 1 states the decision structure this analysis implies for a deployer or evaluator who observes a rise in ASR under degradation.

The algorithm's step 1 threshold is a convention and not a test; the per-level standard error near 0.5 for 770 items is about 0.018 by binomial arithmetic, and three of those is about 0.05.

8. Implications for visual-text compression

The anchor's practical concern is that visual-text compression makes a degraded rendering a routine operation. That concern survives the analysis above, but its shape depends on the account.

In a compression pipeline, the deployer chooses the resolution. The adversary controls the text that gets rendered but not the DPI, unless the pipeline adapts resolution to content length. DeepSeek-OCR's precision depends on the compression ratio (Wei, Sun and Li, 2025), and the dependence on language priors that Liang et al. (2026) document grows as visual tokens shrink. A pipeline that compresses aggressively to fit a long context into a fixed budget can therefore push some rendered content into the region where Song et al. see elevated ASR, without any adversarial choice of resolution. Whether rendered context at compression ratios used in practice falls in that region for a given model is not known; the anchor's sweep uses short instructions cropped to their bounding box, not long rendered documents.

The account matters for the fix. If the effect is contention, serialising transcription and audit is a reasonable general remedy, and its cost - Song et al. report output doubling - is the price. If it is coverage, the remedy is to include harmful content rendered by the deployed pipeline, at the deployed resolutions, in safety training, and to expect that each new rendering scheme reopens the gap; the past-tense result suggests targeted data can close a specific gap (Andriushchenko and Flammarion, 2024). Coverage also implies a simpler architectural option for compression pipelines in particular: when the system itself renders text to save tokens, the text exists before it becomes an image, and moderation can be applied to it there. A text moderator applied before rendering, or a harm detector on outputs (Pi et al., 2024), puts the check at a point the degradation does not reach. A multimodal guard such as Llama Guard 3 Vision (Chi et al., 2024) sees the rendered image, so it would itself need testing at the deployed resolutions. That option is not available when the image arrives from outside, which is the setting of typographic injection in agents (Balakrishnan, Mendapara and Garg, 2026) and of the robustness literature on corrupted images (Hendrycks and

Table 3. Predicted change in ACZ attack success under six interventions, by account. "Down", "up" and "no change" are relative to the unmodified ACZ condition; "-" means the account makes no specific prediction. These are the author's derivations from each account as stated in Section 5, not measured values.

Intervention	Contention	Latency	Coverage	Measurement
Transcribe, then answer, with no audit instruction	down	down	down	-
Audit instruction, with no transcription	small change	small change	down	down
Benign secondary task added to a sharp render	up	no change	no change	-
Safety fine-tuning on one degradation, tested on another	down, transfers	down	down, does not transfer	-
Better reading of degraded text, no safety data	down	down	up, peak moves to lower DPI	-
Graded or ground-truth scoring of successes	no change	no change	no change	down

Algorithm 1. What a degradation-induced ASR rise licenses.

Input: ASR at a sweep of image qualities d ; text ASR for the same attack; the prompt template, judge protocol and decoding settings. 1. If the peak exceeds the clean-render ASR by less than about three standard errors of a per-level estimate (for 770 items, about 0.05): report no established degradation effect; stop. 2. Rerun the sweep without the attack template. If the effect largely disappears: report a template-degradation interaction, not a property of degraded input as such. 3. Score successes with a graded or ground-truth evaluator. If ACZ successes are much less informative than clean-render successes: report the harm as overstated by the binary rate. 4. Add a load-only condition (sharp render plus a benign secondary task). If ASR rises: contention is supported. 5. Supply the transcription as text alongside the degraded image. If ASR falls to the clean-render level: latency or coverage is supported; read a late-layer harmfulness probe to separate them. 6. Fine-tune on harmful requests at one degradation type and test on another. If the fix does not transfer: coverage is supported, and any deployed patch must be re-tested for every degradation in use. 7. Choose the defence by the supported account: contention → serialise reading and auditing (e.g. SCO); latency → late-layer or output-side harm detection; coverage → safety data covering the deployed rendering pipeline, or moderation applied to text before rendering; measurement → fix the metric before fixing the model.
--

Dietterich, 2019; Qiu et al., 2022).

None of this makes the efficiency case for rendered text wrong. It means that the safety evaluation of a compression scheme should be run on the scheme's own resolutions and content lengths, against the model's text ASR for the same content, and with the template held fixed.

9. Measurements that would settle the open parts

Five measurements would decide most of what Sections 5 to 7 leave open. Each uses data or procedures that already exist.

1. **A load-only condition.** Present sharp renders of the attack set together with a demanding benign secondary task in the same image or the same turn, matched in token count to the ACZ condition. A rise in ASR supports contention; no change refutes it as the main driver.
2. **A transcription-supplied condition.** Present the ACZ image together with its ground-truth transcription as text, with no audit instruction. If ASR falls to the clean-render level, the model refuses once harmful content is available early,

which supports latency or coverage over contention. Run the reverse - supply a transcription with the harmful words masked - to test whether those words are what the model fails to resolve.

3. **A reading-only fine-tune.** Fine-tune a model to read low-resolution text better, with no safety data, and repeat the sweep. Contention predicts a smaller comfort zone; coverage predicts that it moves to lower resolution and may peak higher, as bijection learning and custom ciphers do in text.
4. **A transfer test for safety data.** Add harmful requests rendered at one degradation to safety fine-tuning, in the manner of VGuard, and test on another. A fix that does not transfer is the signature of a coverage gap.
5. **Graded scoring with the request visible to the judge.** Re-score ACZ and clean-render successes with a StrongREJECT-style rubric or a ground-truth task, and report the informativeness of successes by resolution. This decides how much of the comfort zone is harmful content and how much is affirmative compliance with a partly read request.

Two reporting changes would make the existing literature more usable. Studies should report text ASR, clean-render ASR and degraded ASR for the same items under one judge, and should report the severity of each perturbation rather than a single selected value. And a study that names a mechanism should run at least one condition in which that mechanism and its nearest rival predict different outcomes.

10. Objections

"Song et al. show the effect on ten models, including frontier ones; Balakrishnan et al. used four. The larger study should win." The two studies measure different conditional quantities, and the larger one's effect depends heavily on its template (Section 3.4). Within Song et al.'s own data, three of ten models do not show the inverted U. The question is not which study is right but under what conditions the effect appears, and the published data identify one condition - the template - that strongly modulates it.

"The external OCR result does not matter; the model can read the text, as its own transcription shows." It matters for the mechanism. If the model reads the whole request, a failure to refuse is a failure of safety. If it reads the template and partly reads the harmful phrase, a high transcription score can come from priors on the template, and the failure may be one of comprehension, which changes what fix is needed. Song et al. report word accuracy over the whole instruction, so the data do not distinguish these.

"Cognitive overload and mismatched generalisation are the same thing described at different levels." They make opposite predictions about the effect of better reading (Table 3, row 5) and about the effect of a load-only manipulation (row 3). Two accounts that disagree about the outcome of an experiment are not the same account.

"Whatever the mechanism, SCO works, and that is what practitioners need." SCO works on the conditions it was tested on. Whether it keeps working depends on the mechanism: under coverage, an attacker who places the harmful content in a form that survives transcription - for example a request whose harmful part is only apparent after a later decoding step, as in DRA or bijection attacks - may not be stopped by auditing the transcription. Under contention, it should keep working wherever reading is the bottleneck. A practitioner choosing between SCO, a pre-render text filter and targeted safety data needs to know which.

"The anchor's probe directly shows the safety computation being crowded out." It shows a text-trained probe separating ACZ inputs later in depth than sharp ones. That is consistent with crowding, with slower decoding of degraded glyphs, and with a larger cross-modal shift at low resolution. A probe cannot distinguish them without a control for each (Section 6).

11. What is not known

It is not known whether the comfort zone appears without a jailbreak template at a size that matters; the only template-free numbers are two models in one table. It is not known whether successes in the comfort zone are as informative to an attacker as successes on sharp renders. It is not known which words of the harmful request models read correctly at ACZ resolutions. It is not known whether adding load without changing the input raises ASR on rendered text, whether improving reading without adding safety data shrinks or shifts the comfort zone, or whether safety data for one degradation transfers to another. It is not known whether the resolutions used by deployed visual-text compression pipelines fall in the comfort zone of any production model. No source reviewed here reports the decoding temperature of the target models in the comfort-zone experiments, or a repeated-sample estimate of per-item reliability.

12. Conclusion

Degrading a rendered harmful request does something to an MLLM's safety behaviour. Two independent studies disagree about where ASR leaves zero, but agree that by about 6 to 11 pixels of glyph size it has risen to a peak or a plateau: eight of the ten models in one study peak there, and the other study's curves level off there, at heights that range from under 8 percent for GPT-4o to about 75 percent for Mistral. Above that, one finds a peak that falls back and the other a plateau, and the two studies also differ in prompt template, scoring protocol and models; the published data show that the template alone multiplies the comfort-zone effect several times over. The inverted U that gives the comfort zone its name is what any system produces when it can act on a request at a lower image quality than it can recognise that the request should be refused, and resource contention is only one account of why that might happen. Refusal decided before late-arriving features and safety training that never covered degraded renders produce the same curve, a success metric that counts compliance rather than harm changes its height, and all three are consistent with the same defence. Which of them holds decides whether a deployer should serialise reading and checking, move the check to where the text still exists, or train on the renderings it actually uses. The experiments that would decide it are cheap, and none has been published for rendered text.

References

- Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, 136(4), 569-576. doi:10.1037/0096-3445.136.4.569.
- Andriushchenko, M., & Flammarion, N. (2024). Does Refusal Training in LLMs Generalize to the Past Tense?. arXiv:2407.11969. doi:10.48550/arXiv.2407.11969. [PREPRINT]

- Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., et al. (2024). Refusal in Language Models Is Mediated by a Single Direction. arXiv:2406.11717. doi:10.48550/arXiv.2406.11717. [PREPRINT]
- Azulay, A., Dubinski, J., Li, Z., Mittal, A., & Gandelsman, Y. (2026). Jailbreaking Vision-Language Models Through the Visual Modality. arXiv:2605.00583. doi:10.48550/arXiv.2605.00583. [PREPRINT]
- Bailey, L., Ong, E., Russell, S., & Emmons, S. (2023). Image Hijacks: Adversarial Images can Control Generative Models at Runtime. ICML 2024. arXiv:2309.00236. doi:10.48550/arXiv.2309.00236.
- Balakrishnan, R., Mendapara, S., & Garg, A. (2026). Reading Between the Pixels: Linking Text-Image Embedding Alignment to Typographic Attack Success on Vision-Language Models. arXiv:2604.12371. doi:10.48550/arXiv.2604.12371. [PREPRINT]
- Belinkov, Y. (2021). Probing Classifiers: Promises, Shortcomings, and Advances. Computational Linguistics. arXiv:2102.12452. doi:10.48550/arXiv.2102.12452.
- Cai, H., Arunasalam, A., Lin, L. Y., Bianchi, A., & Celik, Z. B. (2024). Rethinking How to Evaluate Language Model Jailbreak. arXiv:2404.06407. doi:10.48550/arXiv.2404.06407. [PREPRINT]
- Chao, P., Debenedetti, E., Robey, A., Andriushchenko, M., Croce, F., Schwag, V., et al. (2024). JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. NeurIPS 2024 Datasets and Benchmarks Track. arXiv:2404.01318. doi:10.48550/arXiv.2404.01318.
- Cheng, J., Liu, Y., Zhang, X., Fei, Y., Hong, W., Lyu, R., et al. (2025). Glyph: Scaling Context Windows via Visual-Text Compression. arXiv:2510.17800. doi:10.48550/arXiv.2510.17800. [PREPRINT]
- Chi, J., Karn, U., Zhan, H., Smith, E., Rando, J., Zhang, Y., et al. (2024). Llama Guard 3 Vision: Safeguarding Human-AI Image Understanding Conversations. arXiv:2411.10414. doi:10.48550/arXiv.2411.10414. [PREPRINT]
- Deng, Y., Zhang, W., Pan, S. J., & Bing, L. (2023). Multilingual Jailbreak Challenges in Large Language Models. ICLR 2024. arXiv:2310.06474. doi:10.48550/arXiv.2310.06474.
- Downer, G., Craven, S., Ruck, D., & Thomas, J. (2025). Text2VLM: Adapting Text-Only Datasets to Evaluate Alignment Training in Visual Language Models. arXiv:2507.20704. doi:10.48550/arXiv.2507.20704.
- Gong, Y., Ran, D., Liu, J., Wang, C., Cong, T., Wang, A., et al. (2023). FigStep: Jailbreaking Large Vision-Language Models via Typographic Visual Prompts. AAAI 2025. arXiv:2311.05608. doi:10.48550/arXiv.2311.05608.
- Gou, Y., Chen, K., Liu, Z., Hong, L., Xu, H., Li, Z., et al. (2024). Eyes Closed, Safety On: Protecting Multimodal LLMs via Image-to-Text Transformation. ECCV 2024. arXiv:2403.09572. doi:10.48550/arXiv.2403.09572.
- Guan, M. Y., Joglekar, M., Wallace, E., Jain, S., Barak, B., Helyar, A., et al. (2024). Deliberative Alignment: Reasoning Enables Safer Language Models. arXiv:2412.16339. doi:10.48550/arXiv.2412.16339. [PREPRINT]
- Handa, D., Zhang, Z., Saeidi, A., Kumbhar, S., Uddin, M. N., RRV, A., et al. (2024). When "Competency" in Reasoning Opens the Door to Vulnerability: Jailbreaking LLMs via Novel Complex Ciphers. arXiv:2402.10601. doi:10.48550/arXiv.2402.10601. [PREPRINT]
- Hendrycks, D., & Dietterich, T. (2019). Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. ICLR 2019. arXiv:1903.12261. doi:10.48550/arXiv.1903.12261.
- Hewitt, J., & Liang, P. (2019). Designing and Interpreting Probes with Control Tasks. EMNLP 2019. arXiv:1909.03368. doi:10.48550/arXiv.1909.03368.
- Huang, B. R. Y., Li, M., & Tang, L. (2024). Endless Jailbreaks with Bijection Learning. arXiv:2410.01294. doi:10.48550/arXiv.2410.01294. [PREPRINT]
- Hughes, J., Price, S., Lynch, A., Schaeffer, R., Barez, F., Koyejo, S., et al. (2024). Best-of-N Jailbreaking. arXiv:2412.03556. doi:10.48550/arXiv.2412.03556. [PREPRINT]
- Jiang, F., Xu, Z., Niu, L., Xiang, Z., Ramasubramanian, B., Li, B., et al. (2024). ArtPrompt: ASCII Art-based Jailbreak Attacks against Aligned LLMs. ACL 2024. arXiv:2402.11753. doi:10.48550/arXiv.2402.11753.
- Li, L., Dong, B., Wang, R., Hu, X., Zuo, W., Lin, D., et al. (2024b). SALAD-Bench: A Hierarchical and Comprehensive Safety Benchmark for Large Language Models. Findings of ACL 2024. arXiv:2402.05044. doi:10.48550/arXiv.2402.05044.
- Li, S., Yao, L., Zhang, L., & Li, Y. (2024c). Safety Layers in Aligned Large Language Models: The Key to LLM Security. ICLR 2025. arXiv:2408.17003. doi:10.48550/arXiv.2408.17003.
- Li, X., Zhou, H., Wang, R., Zhou, T., Cheng, M., & Hsieh, C. J. (2024d). MOSSBench: Is Your Multimodal Language Model Oversensitive to Safe Queries?. arXiv:2406.17806. doi:10.48550/arXiv.2406.17806. [PREPRINT]
- Li, Y., Guo, H., Zhou, K., Zhao, W. X., & Wen, J. R. (2024a). Images are Achilles' Heel of Alignment: Exploiting Visual Vulnerabilities for Jailbreaking Multimodal Large Language Models. ECCV 2024. arXiv:2403.09792. doi:10.48550/arXiv.2403.09792.
- Li, Y., Lan, Z., & Zhou, J. (2025). Text or Pixels? It Takes Half: On the Token Efficiency of Visual Text Inputs in Multimodal LLMs. Findings of EMNLP 2025. arXiv:2510.18279. doi:10.48550/arXiv.2510.18279.
- Liang, Y., Ying, R., Li, B., Li, H., Yan, K., Li, Q., et al. (2026). Visual Merit or Linguistic Crutch? A Close Look at DeepSeek-OCR. arXiv:2601.03714. doi:10.48550/arXiv.2601.03714. [PREPRINT]
- Liu, Q., Shang, C., Liu, L., Pappas, N., Ma, J., John, N. A., et al. (2024b). Unraveling and Mitigating Safety Alignment Degradation of Vision-Language Models. arXiv:2410.09047. doi:10.48550/arXiv.2410.09047. [PREPRINT]
- Liu, T., Zhang, Y., Zhao, Z., Dong, Y., Meng, G., & Chen, K. (2024a). Making Them Ask and Answer: Jailbreaking Large Language Models in Few Queries via Disguise and Reconstruction. USENIX Security 2024. arXiv:2402.18104. doi:10.48550/arXiv.2402.18104.
- Liu, X., Zhu, Y., Gu, J., Lan, Y., Yang, C., & Qiao, Y. (2023a). MM-SafetyBench: A Benchmark for Safety Evaluation of Multimodal Large Language Models. arXiv:2311.17600. doi:10.48550/arXiv.2311.17600.
- Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., et al. (2023b). OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models. arXiv:2305.07895. doi:10.48550/arXiv.2305.07895. [PREPRINT]
- Luo, W., Ma, S., Liu, X., Guo, X., & Xiao, C. (2024). JailBreakV: A Benchmark for Assessing the Robustness of MultiModal Large Language Models against Jailbreak Attacks. arXiv:2404.03027. doi:10.48550/arXiv.2404.03027. [PREPRINT]
- Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., et al. (2024). HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. ICML 2024. arXiv:2402.04249. doi:10.48550/arXiv.2402.04249.
- Meyer, A., Frederick, S., Burnham, T. C., Guevara Pinto, J. D., Boyer, T. W., Ball, L. J., Pennycook, G., Ackerman, R., et al. (2015). Disfluent fonts don't help people solve math problems. Journal of Experimental Psychology: General, 144(2), e16-e30. doi:10.1037/xge0000049.
- Nikolic, K., Sun, L., Zhang, J., & Tramer, F. (2025). The Jailbreak Tax: How Useful are Your Jailbreak Outputs?. arXiv:2504.10694. doi:10.48550/arXiv.2504.10694.
- Pi, R., Han, T., Zhang, J., Xie, Y., Pan, R., Lian, Q., et al. (2024). MLLM-Protector: Ensuring MLLM's Safety without Hurting Performance. arXiv:2401.02906. doi:10.48550/arXiv.2401.02906.
- Qi, X., Huang, K., Panda, A., Henderson, P., Wang, M., & Mittal, P. (2023). Visual Adversarial Examples Jailbreak Aligned Large Language Models. AAAI 2024. arXiv:2306.13213. doi:10.48550/arXiv.2306.13213.
- Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., et al. (2024). Safety Alignment Should Be Made More Than Just a Few Tokens Deep. arXiv:2406.05946. doi:10.48550/arXiv.2406.05946.

- Qiu, J., Zhu, Y., Shi, X., Wenzel, F., Tang, Z., Zhao, D., et al. (2022). Benchmarking Robustness of Multimodal Image-Text Models under Distribution Shift. *arXiv:2212.08044*. doi:10.48550/arXiv.2212.08044.
- Qraitem, M., Tasnim, N., Teterwak, P., Saenko, K., & Plummer, B. A. (2024). Vision-LLMs Can Fool Themselves with Self-Generated Typographic Attacks. *arXiv:2402.00626*. doi:10.48550/arXiv.2402.00626. [PREPRINT]
- Rotger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., & Hovy, D. (2023). XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. *NAACL 2024*. *arXiv:2308.01263*. doi:10.48550/arXiv.2308.01263.
- Rust, P., Lotz, J. F., Bugliarello, E., Salesky, E., de Lhoneux, M., & Elliott, D. (2022). Language Modelling with Pixels. *ICLR 2023*. *arXiv:2207.06991*. doi:10.48550/arXiv.2207.06991.
- Shayegani, E., Dong, Y., & Abu-Ghazaleh, N. (2023). Jailbreak in pieces: Compositional Adversarial Attacks on Multi-Modal Language Models. *arXiv:2307.14539*. doi:10.48550/arXiv.2307.14539.
- Song, Z., Han, B., Wang, Y., & Zhang, C. (2026). Hard to Read, Easy to Jailbreak: How Visual Degradation Bypasses MLLM Safety Alignment. *Findings of ACL 2026*. *arXiv:2605.07250*. doi:10.48550/arXiv.2605.07250.
- Souly, A., Lu, Q., Bowen, D., Trinh, T., Hsieh, E., Pandey, S., et al. (2024). A StrongREJECT for Empty Jailbreaks. *arXiv:2402.10260*. doi:10.48550/arXiv.2402.10260.
- Wang, X., Wang, M., Liu, Y., Schutze, H., & Plank, B. (2025). Refusal Direction is Universal Across Safety-Aligned Languages. *arXiv:2505.17306*. doi:10.48550/arXiv.2505.17306. [PREPRINT]
- Wang, Y., Liu, X., Li, Y., Chen, M., & Xiao, C. (2024a). AdaShield: Safeguarding Multimodal Large Language Models from Structure-based Attack via Adaptive Shield Prompting. *arXiv:2403.09513*. doi:10.48550/arXiv.2403.09513.
- Wattenberg, M., Viegas, F., & Johnson, I. (2016). How to Use t-SNE Effectively. *Distill*. doi:10.23915/distill.00002.
- Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How Does LLM Safety Training Fail?. *arXiv:2307.02483*. doi:10.48550/arXiv.2307.02483.
- Wei, H., Sun, Y., & Li, Y. (2025). DeepSeek-OCR: Contexts Optical Compression. *arXiv:2510.18234*. doi:10.48550/arXiv.2510.18234. [PREPRINT]
- Xiao, Z., Yang, Y., Chen, G., & Chen, Y. (2024). Distract Large Language Models for Automatic Jailbreak Attack. *EMNLP 2024*. *arXiv:2403.08424*. doi:10.48550/arXiv.2403.08424.
- Xie, Y., Yi, J., Shao, J., Curl, J., Lyu, L., Chen, Q., Xie, X., & Wu, F. (2023). Defending ChatGPT against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5, 1486-1496. doi:10.1038/s42256-023-00765-8.
- Xing, S., Guo, L., Hua, H., Lee, S., Li, P., Wang, Y., et al. (2025). Demystifying the Visual Quality Paradox in Multimodal Large Language Models. *arXiv:2506.15645*. doi:10.48550/arXiv.2506.15645. [PREPRINT]
- Xu, N., Wang, F., Zhou, B., Li, B. Z., Xiao, C., & Chen, M. (2023). Cognitive Overload: Jailbreaking Large Language Models with Overloaded Logical Thinking. *arXiv:2311.09827*. doi:10.48550/arXiv.2311.09827.
- Xu, S., Pang, L., Zhu, Y., Shen, H., & Cheng, X. (2024). Cross-Modal Safety Mechanism Transfer in Large Vision-Language Models. *ICLR 2025*. *arXiv:2410.12662*. doi:10.48550/arXiv.2410.12662.
- Yang, Z., Fan, J., Yan, A., Gao, E., Lin, X., Li, T., et al. (2025). Distraction is All You Need for Multimodal Large Language Model Jailbreaking. *CVPR 2025*. *arXiv:2502.10794*. doi:10.48550/arXiv.2502.10794.
- Yong, Z. X., Menghini, C., & Bach, S. H. (2023). Low-Resource Languages Jailbreak GPT-4. *arXiv:2310.02446*. doi:10.48550/arXiv.2310.02446. [PREPRINT]
- Yuan, Y., Jiao, W., Wang, W., Huang, J. T., He, P., Shi, S., et al. (2023). GPT-4 Is Too Smart To Be Safe: Stealthy Chat with LLMs via Cipher. *ICLR 2024*. *arXiv:2308.06463*. doi:10.48550/arXiv.2308.06463.
- Zhao, J., Huang, J., Wu, Z., Bau, D., & Shi, W. (2025). LLMs Encode Harmfulness and Refusal Separately. *arXiv:2507.11878*. doi:10.48550/arXiv.2507.11878. [PREPRINT]
- Zhao, W., Li, Z., Li, Y., Zhang, Y., & Sun, J. (2024). Defending Large Language Models Against Jailbreak Attacks via Layer-specific Editing. *Findings of EMNLP 2024*. *arXiv:2405.18166*. doi:10.48550/arXiv.2405.18166.
- Zheng, W., Zeng, P., Li, Y., Wu, H., Lin, N., Chen, J., et al. (2024). Jailbreaking? One Step Is Enough!. *ACL 2025*. *arXiv:2412.12621*. doi:10.48550/arXiv.2412.12621.
- Zong, Y., Bohdal, O., Yu, T., Yang, Y., & Hospedales, T. (2024). Safety Fine-Tuning at (Almost) No Cost: A Baseline for Vision Large Language Models. *ICML 2024*. *arXiv:2402.02207*. doi:10.48550/arXiv.2402.02207.
- Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv:2307.15043*. doi:10.48550/arXiv.2307.15043. [PREPRINT]