CAN AI HELP REGULATE AI? BENCHMARKING MULTIMODAL MODELS AS ADVERTISING REGULATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI generation is scaling far faster than human regulation. Generative models enable regulated content to be produced at machine speed and near-zero marginal cost, with multimodal advertising generation providing a prominent example, while regulatory review remains constrained by scarce human attention. This growing asymmetry creates a scalable regulation problem: can AI help humans regulate AI-generated content at the scale at which it is produced? We introduce RegulaBench, a benchmark for evaluating multimodal models as advertising pre-regulators—scalable assistants that screen potentially unlawful content, organize legally relevant evidence, and provide grounded regulatory recommendations for subsequent human review. Built from real advertising enforcement cases, RegulaBench links generated advertising images to the factual findings and legal grounds recorded in corresponding administrative penalty decisions. It evaluates three progressively demanding capabilities: violation detection, whether models identify advertisements that trigger regulatory intervention; evidence-to-norm grounding, whether they connect case-specific evidence to the conditions governing applicable legal provisions; and regulatory fidelity, whether their selection and application of legal norms remain aligned with the structure of human regulatory decisions. Our experiments reveal a two-sided picture of AI-assisted regulation. On the one hand, across a substantial share of cases, model judgments are consistent with those of human regulators, suggesting that AI can help extend human regulatory capacity in screening potentially unlawful advertising at scale. On the other hand, this agreement becomes substantially weaker once we move beyond regulatory outcomes to examine how those outcomes are legally justified. In particular, across the provision groups examined, models show a directional tendency to rely more heavily on specific content prohibitions, whereas human enforcement decisions more often invoke broader truthfulness provisions. These findings reveal an important tension in scalable AI-assisted regulation: AI can scale regulation without faithfully reproducing existing patterns of human regulation. Such divergence raises consequential questions about how AI pre-regulators should be integrated into human regulatory systems.

1 INTRODUCTION

Advances in large language models (LLMs) and multimodal models have opened new possibilities for AI-assisted regulatory decision-making. Evaluating models in this role requires examining not only whether they reach legally correct conclusions, but also how they apply the law. Legal systems rely on rules, standards, and principles that differ in their level of abstraction, and regulators' choices among these norms shape how legal frameworks operate in practice. Where several grounds are legally available, a model may reach the same conclusion as a human regulator while relying on a different type of norm. If such differences recur systematically, AI-assisted regulation could reshape patterns of legal application even when the underlying law remains unchanged.

Existing benchmarks have substantially advanced the evaluation of legal knowledge and reasoning. LegalBench examines multiple forms of legal reasoning, while LawBench covers a broad range of Chinese legal tasks Guha et al. (2023); Fei et al. (2024). Yet performance on these tasks does not

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.

reveal how models choose among multiple legally applicable grounds. Correctly identifying a relevant provision, for example, does not show whether a model systematically prefers concrete rules over broader standards or principles, or whether its choices resemble those made in actual enforcement. This distinction matters because recurring differences in norm selection can affect what kinds of conduct receive regulatory attention and how legal obligations are interpreted in practice.

We use **norm-selection preference** to describe a decision-maker's systematic tendency to invoke legal norms at particular levels of abstraction when multiple grounds are legally applicable. This tendency is analytically distinct from violation detection or provision-level correctness: two decisions may agree that conduct is unlawful yet justify that conclusion through different forms of legal reasoning. A mismatch with an enforcement decision may therefore reflect legal error, an alternative permissible ground, or a recurring preference for a particular type of norm. Evaluating such preferences requires moving beyond citation agreement to examine whether differences from human enforcement show a consistent direction across cases and informational conditions.

We study norm-selection preferences in the regulation of image advertisements in China. Advertising law combines specific content prohibitions with broader provisions governing false or misleading claims, so a single advertisement may implicate several norms at different levels of abstraction. The task is also inherently multimodal: regulatory meaning can depend on visual representations, textual claims, and their interaction. Administrative penalty decisions further provide documented legal grounds from actual enforcement, offering an empirical reference for comparing model judgments with human regulatory practice. These features make advertising a useful setting for studying how models select among legally available norms.

We introduce **JurisBench-ImageRecognition**, a benchmark linking multimodal advertising judgments to documented human enforcement decisions. The benchmark uses advertising images generated from real administrative enforcement cases and associates each image with the legal grounds cited in its source decision. Models are asked to determine whether an advertisement violates applicable law and to identify the provisions supporting that judgment. We separate two dimensions of model behavior:

1. **Violation Detection.** We measure whether models identify violations associated with rules, standards, and principles, allowing us to examine whether detection varies with the level of legal abstraction.
2. **Norm Selection.** We compare the legal grounds invoked by models with those recorded in the corresponding enforcement decisions, focusing on both shared grounds and directional shifts across levels of abstraction.

To distinguish norm-selection preferences from information deficits, we further introduce controlled conditions that supplement the same advertisements with factual context, relevant legal provisions, or both. These comparisons test whether observed differences diminish with additional information or persist across informational conditions.

Our preliminary analysis reveals a directional asymmetry within the provision groups examined: models more often shift from broad truthfulness provisions cited in human enforcement decisions toward specific content prohibitions than in the reverse direction. These differences can occur even when models and human regulators agree that an advertisement is unlawful, showing that outcome agreement can conceal systematic differences in legal justification. JurisBench-ImageRecognition therefore extends legal AI evaluation from whether models reach the same conclusions as regulators to how they distribute regulatory judgment across different forms of legal norms. More broadly, the results suggest that AI-assisted regulation may influence the practical balance among rules, standards, and principles through changes in norm application rather than changes in legal text.

2 Related Work

Legal and regulatory benchmarks. LegalBench and LawBench evaluate broad legal-reasoning and Chinese-law capabilities (Guha et al., 2023; Fei et al., 2024), while LAiW, JurisBench, and ALARB expose intermediate stages linking facts, reasoning, outcomes, and legal authorities (Dai et al., 2025; Chen et al., 2026; Abu Shairah et al., 2025). RoD-TAL extends legally grounded evaluation to visual driving-examination questions (Man et al., 2026). Regulatory benchmarks narrow

the task further: AIReg-Bench evaluates compliance under specified EU AI Act provisions (Marino et al., 2025); Rad et al. study article attribution from GDPR enforcement summaries (Rad et al., 2026); and Takamatsu et al. assess advertising risk and applicable law from Japanese administrative and judicial materials (Takamatsu et al., 2025). Most closely, RegNLI links FDA warning letters, packaging images, misbranding judgments, and regulatory sections for extracted health-product claims (Saha et al., 2026). These studies move from general competence toward institutional decisions, but largely evaluate answer or citation accuracy rather than whether models systematically select different types of legally available norms from human decision-makers.

Advertising understanding and compliance. Advertising research first established that interpreting an ad requires more than recognizing objects. PittAds annotates intended actions and communicative rationales (Hussain et al., 2017); ADVISE models symbolism and external knowledge (Ye & Kovashka, 2018); and later work combines scene text, captions, visual cues, and entity knowledge (Kalra et al., 2020; Jia et al., 2023). Recent benchmarks also examine atypical persuasive imagery and aesthetic effects (Malakouti et al., 2025; Ji et al., 2026). Compliance-oriented work adds decision criteria: M-PACE evaluates brand alignment and content safety, while CompAgent checks visual content against supplied policies (Verma et al., 2025; Ghosh et al., 2026). RAVEN, RAVEN++, and BLM-Guard study advertisement-video violations and their localization (Ji et al., 2025b;a; Yang et al., 2026), and Levi et al. compare multimodal models with professional moderators (Levi et al., 2025). Gui et al. move closer to law by evaluating sponsored-post identification and explanations under Dutch advertising codes (Gui et al., 2025). Yet interpretation, persuasion, brand safety, platform policy, and disclosure do not by themselves test provision-level statutory judgment across diverse advertising violations.

Positioning our work. Our benchmark joins these strands by linking image-level advertising judgments to legal grounds documented in Chinese administrative enforcement. Beyond violation detection, it compares recurring model and human patterns of norm selection. Rules, standards, and principles differ in how they structure legal application (Kaplow, 1992; Dworkin, 1967), and AI-assisted moderation may change the practical role of rule-like language (Keydar et al., 2026). We therefore treat norm type as a pre-specified analytic category and examine directional shifts across levels of abstraction, rather than assuming that every departure from an enforcement record is an error. The comparison uses image-supported, case-derived institutional references; it is not a same-image study in which human regulators and models receive identical evidence.

3 BENCHMARK

3.1 BENCHMARK OVERVIEW

RegulaBench evaluates multimodal models as advertising pre-regulators. It contains 6,011 generated advertisements: 2,549 case-derived images and 3,462 rule-derived images. Case-derived images reconstruct advertising scenarios described in real administrative penalty decisions publicly released by China's State Administration for Market Regulation (SAMR), preserving links to the documented facts and cited legal grounds. Rule-derived images instantiate statutory and regulatory provisions, extending coverage beyond observed enforcement cases. RegulaBench records the provenance, generation prompt, and associated legal materials for every image. This paper evaluates a 600-image set centered on the case-derived track rather than the entire corpus. Figure 1 summarizes the four-stage evaluation framework.

3.2 DATASET CONSTRUCTION

For the case-derived track, we convert only the decisions' factual descriptions—including textual claims, visual elements, product categories, and presentation contexts—into image-generation prompts; the cited provisions never enter the prompts. A persistent identifier links each image to its source decision, prompt, and cited provisions. For the rule-derived track, individual statutory or regulatory provisions are converted into advertising scenarios designed to instantiate the regulated conduct. The present evaluation set contains 400 positive images from the case-derived track and 200 human-reviewed images with no image-evident violation.

3.3 Evaluation Tasks

In the baseline setting, the model receives one advertising image without source-case facts or reference legal grounds. It returns a structured response containing a violation judgment, law names, specific provisions, the content considered unlawful, and a brief justification. Specifically, **1) Violation Detection** asks whether the advertisement contains an identifiable violation; **2) Law Identification** asks which laws or regulations apply to that violation; and **3) Provision Grounding** asks for the supporting law, article, and, where applicable, paragraph or item. All three outputs are produced in one response. The unlawful-content and justification fields support auditing but are not scored.

3.4 Evaluation Design

The evaluation has three components. **1) Baseline Evaluation.** Violation Detection is measured on all 600 images using accuracy, recall, and specificity, while Law Identification and Provision Grounding are evaluated conditionally on positive images with valid affirmative judgments. Laws use strict set exact match and partial coverage; provisions use overlap and All-Gold-Provisions-Hit, which permits extra citations and is therefore not exact match. For regulatory fidelity, provisions are coded before comparison as *Rules*, *Standards*, *Principles*, or *Mixed*: Rules enumerate duties or prohibitions, Standards require open-textured assessment, Principles express broad public-order or social-value commitments, and heterogeneous provisions remain Mixed. We compare incorrect-citation concentration and norm-type distributions with the case-derived references. **2) Scene-Description Intervention.** Audited descriptions of visible layout and literal text are added to selected image-only false negatives, without legal conclusions or statutes, and we measure repair, regressions, and invalid outputs. **3) Known-Law Intervention.** When a model identifies a violation but misses the reference law and provisions, it receives only the correct law name, without articles or additional facts, to test whether this information improves provision grounding.

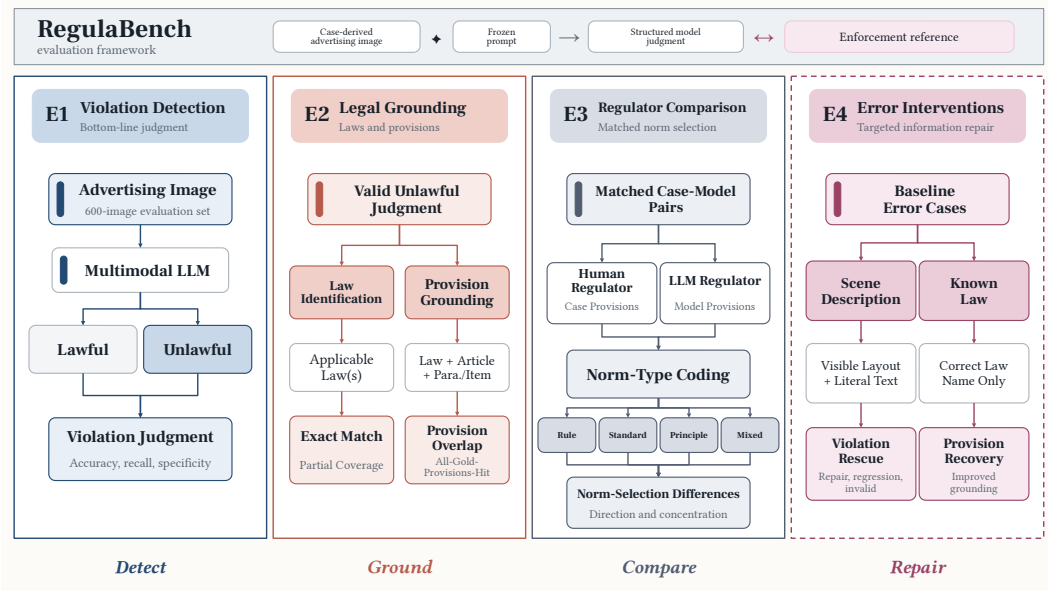

Figure 1: RegulaBench evaluation framework. The benchmark separates binary violation detection, law and provision grounding, comparison of human and LLM norm selection, and two targeted information interventions.

4 Experiments

4.1 From Images to Violation Judgments

We first evaluate whether a model can recover the bottom-line regulatory judgment from the image alone. The current frozen set contains 400 image-evident violation cases and 200 human-reviewed

images with no image-evident violation. All 200 reviewed negative images met the eligibility criterion and were retained. Violation Judgment Accuracy over all eligible images is the primary outcome. We additionally report violation recall, specificity, explicit false negatives and false positives, and invalid binary judgments. An API-successful answer is a valid binary judgment only if its parsed output contains the canonical affirmative or negative value in the required field. Invalid model answers remain in the metric denominators and count as incorrect, whereas API transport failures are reported as incomplete coverage rather than model errors. Models with incomplete coverage are reported provisionally over their currently successful outputs and excluded from complete-coverage rankings.

Table 1: Image-level violation-judgment performance. Accuracy uses all 600 images, recall the 400 violation cases, and specificity the 200 reviewed non-violation images. Full counts and intervals appear in Appendix A.

Family	Model	Accuracy	Recall	Specificity
Qwen	Qwen3-VL-Plus	78.7	<u>99.8</u>	36.5
Qwen	Qwen3-VL-Flash	82.2	96.2	54.0
DeepSeek	DeepSeek-V4.1-Flash	74.5	98.0	27.5
OpenAI	GPT-4o	**88.0**	90.5	**83.0**
OpenAI	GPT-4o-mini	78.8	95.2	46.0
Google	Gemini-2.5-Pro	70.5	**100.0**	11.5
Google	Gemini-2.5-Flash	74.7	96.0	32.0
Anthropic	Claude-Sonnet-4.5	77.7†	89.0†	53.3†
Anthropic	Claude-Haiku-4.5	63.7†	81.5†	25.0†
xAI	Grok-4.6	<u>84.7</u>	98.8	56.5
GLM	GLM-4.6V	82.5	94.0	<u>59.5</u>

Bold and <u>underlining</u> mark the best and second-best complete-coverage results. †Provisional rows are excluded from ranking.

The results in Table 1 show that high violation recall does not by itself imply reliable screening. Among the nine models with complete API coverage, GPT-4o obtains the highest overall accuracy (88.0%) and specificity (83.0%), although its violation recall (90.5%) is lower than that of several models. At the opposite extreme, Gemini-2.5-Pro identifies all 400 violation cases but correctly rejects only 11.5% of the reviewed negative images, producing 177 explicit false positives. Grok-4.6 combines 98.8% recall with 56.5% specificity, while GLM-4.6V and Qwen3-VL-Flash attain similar overall accuracy (82.5% and 82.2%) through different recall–specificity trade-offs. More broadly, most models favor sensitivity over specificity, which would increase the downstream human-review burden of an AI pre-regulator. The Anthropic rows provide preliminary results over their currently successful outputs and are excluded from the complete-coverage ranking.

For the positive-case judgment chain in Sections 4.2 and 4.2.2, we retain the 400 violation cases as a common denominator. We therefore do not connect the all-image accuracy reported here to later exact-match rates as though their denominators were identical.

4.2 From Violation Judgments to Legal Bases and Provisions

4.2.1 Applicable Law Names

We next assess whether the model identifies the applicable law names once it has correctly identified an image as unlawful. This is therefore a conditional analysis: each model's denominator is its own set of correctly detected violation images, rather than all 400 positive images. Law-name exact match requires the predicted law set to equal the image-supported reference set, with neither omissions nor additions.

Table 2 separates this outcome by the number of distinct laws in the reference set. For single-law images, we report strict exact match and over-citation. For two- and three-law images, the table separately reports strict complete matching and cases where the model names exactly one reference law. This distinction keeps a partial legal basis from being conflated with a complete match; the

complete $0/k$, $1/k$, …, k/k breakdown is preserved in the underlying audit data, while the $\geq$4-law cases are not compared here because there are only five.

Table 2: Conditional law-name matching by reference-law count. Each n contains images the model correctly judged unlawful; Exact requires the full reference set with no additions, whereas $1/k$ hit means exactly one reference law is named regardless of additions.

	One reference law (277 images)			Two reference laws (95 images)			Three reference laws (23 images)		
Model	n	Exact	Over-cite	n	Exact	$1/2$ hit	n	Exact	$1/3$ hit
Qwen3-VL-Plus	276	<u>72.1%</u>	27.9%	95	0.0%	78.9%	23	0.0%	91.3%
Qwen3-VL-Flash	265	**72.8%**	27.2%	93	0.0%	77.4%	22	0.0%	90.9%
DeepSeek-V4.1-Flash	270	41.9%	58.1%	94	**35.1%**	48.9%	23	**4.3%**	52.2%
GPT-4o	250	**72.8%**	27.2%	87	0.0%	78.2%	20	0.0%	90.0%
GPT-4o-mini	261	72.0%	28.0%	93	0.0%	79.6%	22	0.0%	90.9%
Gemini-2.5-Pro	277	65.7%	34.3%	95	5.3%	76.8%	23	0.0%	91.3%
Gemini-2.5-Flash	266	50.4%	49.6%	91	18.7%	60.4%	22	0.0%	72.7%
Claude-Sonnet-4.5	248	71.8%	28.2%	82	0.0%	78.0%	22	0.0%	90.9%
Claude-Haiku-4.5	228	11.0%	89.0%	73	<u>23.3%</u>	72.6%	20	0.0%	70.0%
Grok-4.6	272	<u>72.1%</u>	27.9%	95	2.1%	78.9%	23	0.0%	82.6%
GLM-4.6V	257	70.8%	29.2%	92	3.3%	77.2%	22	0.0%	90.9%

Bold and <u>underlining</u> mark the highest and second-highest Exact rates within each stratum; no runner-up is marked when all remaining values are zero.

The single-law results should not be read as a uniform model ranking. Excluding Claude-Haiku-4.5, strict exact match ranges from 41.9% to 72.8%, with Qwen3-VL-Flash and GPT-4o tied at 72.8%. However, the gap between naming the gold law and exact matching is substantial for DeepSeek-V4.1-Flash (86.3% versus 41.9%), Gemini-2.5-Flash (83.8% versus 50.4%), and Claude-Haiku-4.5 (78.9% versus 11.0%). Together with their over-citation rates (58.1%, 49.6%, and 89.0%, respectively), these gaps identify a distinct addition-error pattern rather than a simple failure to name the gold law.

The more consequential pattern is the collapse of complete matching on multi-law images. On the 95 two-law images, Qwen3-VL-Plus, Qwen3-VL-Flash, GPT-4o, GPT-4o-mini, and Claude-Sonnet-4.5 have zero strict complete matches, although each names exactly one of the two reference laws in at least 77.4% of its conditionally retained cases. DeepSeek-V4.1-Flash is the highest two-law exact matcher (35.1%), followed by Claude-Haiku-4.5 (23.3%) and Gemini-2.5-Flash (18.7%), but this advantage coincides with substantially higher over-citation. On the 23 three-law images, only DeepSeek-V4.1-Flash records any strict complete match (4.3%); the other ten models record none. Yet ten models name exactly one of the three reference laws in 70.0%–91.3% of their retained cases. Thus, for multi-law images, the principal observed failure is incomplete legal coverage rather than an absence of any relevant legal basis.

These are conditional rather than end-to-end results, so the varying n values must be retained when comparing models. The three-law stratum is also small ($n \leq 23$ per model), and the five cases with four or more laws are not used for model comparisons. We therefore treat the table as evidence of a robust error pattern across the available cases, not as a significance-tested ranking.

4.2.2 Specific Provisions

We then retain the same conditional entry point—a valid affirmative violation judgment—and ask whether the response identifies the image-supported specific provisions. Each gold provision comprises a law, article, and any annotated paragraph or item. A hit requires all of these components and an unambiguous association with the cited law. This is a provision-overlap analysis: "All" means that every gold provision is hit, without separately penalizing additional provisions cited by the model. It is therefore not provision-set exact match. For example, if the gold set is $\{p_1, p_2\}$ and the model cites $\{p_1, p_2, p_3\}$, all-gold-provisions-hit equals one, whereas provision-set exact match equals zero. The current table reports the former; the latter requires canonical parsing of every additional predicted provision.

Provision matching is substantially weaker than law-name matching. For single-provision images, DeepSeek-V4.1-Flash has the highest all-hit rate (39.8%), followed by Qwen3-VL-Plus (29.8%) and Gemini-2.5-Flash (23.2%). For the 129 two-provision images, the best all-hit rate is only 5.5% (DeepSeek-V4.1-Flash); for the 47 three-provision images, it is 2.1% (also DeepSeek-V4.1-Flash).

Partial overlap persists—for example, DeepSeek-V4.1-Flash overlaps with one of two provisions in 29.1% of its retained cases and with one or two of three provisions in 51.1%—but almost never reaches complete coverage.

Table 3: Conditional provision overlap by gold-provision count. A hit requires the gold law, article, and any annotated paragraph/item; All means every gold provision is hit. The ≥4 stratum (53 images) is omitted.

	One gold provision (171 images)		Two gold provisions (129 images)			Three gold provisions (47 images)			
Model	n	All	n	All	$1/2$	n	All	$2/3$	$1/3$
Qwen3-VL-Plus	171	<u>29.8%</u>	128	0.0%	35.2%	47	0.0%	4.3%	38.3%
Qwen3-VL-Flash	160	4.4%	126	0.0%	3.2%	46	0.0%	0.0%	6.5%
DeepSeek-V4.1-Flash	166	**39.8%**	127	**5.5%**	29.1%	47	**2.1%**	21.3%	29.8%
GPT-4o	149	7.4%	119	0.0%	11.8%	45	0.0%	0.0%	8.9%
GPT-4o-mini	159	0.6%	123	0.0%	1.6%	47	0.0%	0.0%	0.0%
Gemini-2.5-Pro	171	14.0%	129	1.6%	14.7%	47	0.0%	4.3%	21.3%
Gemini-2.5-Flash	164	23.2%	128	<u>3.1%</u>	21.9%	41	0.0%	2.4%	46.3%
Claude-Sonnet-4.5	150	14.0%	121	0.0%	17.4%	41	0.0%	0.0%	24.4%
Claude-Haiku-4.5	138	8.0%	103	1.0%	14.6%	41	0.0%	2.4%	24.4%
Grok-4.6	167	14.4%	128	0.0%	14.1%	47	0.0%	0.0%	17.0%
GLM-4.6V	154	11.7%	123	0.0%	14.6%	46	0.0%	0.0%	10.9%

Bold and <u>underlining</u> mark the highest and second-highest All rates within each stratum; no runner-up is marked when all remaining values are zero.

The multi-provision rows are descriptive: their denominators differ by model because of the conditional entry point, and the three-provision stratum contains at most 47 retained cases per model.

4.3 Where Citation Grounding Breaks: Rules, Standards, and Principles

The provision analysis establishes that a correct violation judgment often fails to yield a complete legal basis. We next ask what the models cite instead. We code each image-supported reference provision as a *Rule*, *Standard*, or *Principle*: Rules specify enumerated prohibitions or procedural duties; Standards depend on truthfulness, misleadingness, efficacy, quality, safety, or other open-textured assessments; and Principles express general public-order or social-value commitments. The coding is fixed from the local provision text and reference labels before comparing it with model outputs.

This analysis remains conditional on a valid affirmative violation judgment. Because model responses generally name an article without a paragraph or item, we compare at the law–article level. A cited article is counted as wrong only when its law–article pair is absent from that image's reference set. A reference article named without the required paragraph or item is instead treated as a granularity-incomplete citation, not as a wrong article. We do not force internally heterogeneous articles into a single type: they form the separate *Mixed* column in Table 4.

Table 4: Norm-type profiles of wrong model law–article citations and corresponding human reference provisions.

Source	n	Rule	Standard	Principle	Mixed
Human reference provisions	8,554	62.6%	35.3%	2.1%	0.0%
Model wrong article citations	4,646	54.0%	27.9%	0.2%	17.8%

Percentages describe citation-type distributions, not provision accuracy. The 525 citations without a reliable law–article pairing are excluded.

The central pattern is concentration rather than a uniform shift toward one of the three norm types. Of the 4,646 identifiable wrong article citations, 4,163 (89.6%) invoke the *Advertising Law*, compared with 4,802 of 8,554 (56.1%) reference-provision appearances in the same conditionally retained cases. Four articles account for 63.2% of all wrong citations: Articles 17 (954 citations), 9 (689), 28 (688), and 16 (603) of the *Advertising Law*. Article 9 is retained as Mixed because its locally available provisions span multiple norm types; treating it as a single Rule or Standard would manufacture a taxonomic result.

Table 4 also shows that Principles are nearly absent from model errors (0.2%) but occur in 2.1% of the corresponding human reference provisions. After keeping Mixed articles separate, the observable error profile is therefore best characterized as repeated reliance on a small set of familiar, mostly rule- or standard-like advertising articles, coupled with little use of principle-based grounds. This is a descriptive comparison of model citation behavior with case-derived reference patterns, not evidence that norm type causes an error or that the reference distribution represents the full population of advertising enforcement decisions.

4.4 Can Additional Information Repair Regulatory Errors?

4.4.1 Scene-Description Intervention

We next test whether a short textual description of the scene can repair image-only errors. The intervention set contains the 38 gold-unlawful images on which GPT-4o produced a valid negative judgment under the frozen image-only prompt. For each image, we retain the image and frozen regulatory instruction, then append a manually audited description limited to visual layout and literal visible text. The description contains no legal conclusion, statute, or assertion about facts not visible in the image. This is an error-repair study on a selected difficult subset, rather than a new estimate of benchmark-wide recall.

Table 5 separates the post-description affirmative count from restoration of each model's own baseline false negatives. The latter comparison uses only valid binary judgments; invalid outputs are reported separately rather than scored as either class.

Table 5: Adding a non-legal scene description to the 38-image GPT-4o error cohort.

Model	Scene: Yes	Restored own FNs	Yes→No	Invalid
Qwen3-VL-Plus	37/38	0/1	0	0
Qwen3-VL-Flash	38/38	8/8	0	0
DeepSeek-V4.1-Flash	33/38	1/2	1	2
GPT-4o	22/38	22/38	0	0
GPT-4o-mini	35/38	7/9	1	0
Gemini-2.5-Pro	38/38	–	0	0
Gemini-2.5-Flash	38/38	–	0	0
Claude-Sonnet-4.5	30/38	6/8	4	1
Claude-Haiku-4.5	15/38	0/2	4	17
Grok-4.6	32/38	0/4	2	0
GLM-4.6V	28/38	8/15	0	3

"Restored own FNs" means valid baseline No → valid post-description Yes; "Invalid" means the required binary label is missing.

Scene information helps some models but not all. It restores 22/38 GPT-4o false negatives and 8/8 evaluable Qwen3-VL-Flash false negatives, whereas Grok restores none of its four and Claude-Haiku returns 17 invalid outputs. Because the cohort was selected from GPT-4o errors and models have different numbers of baseline false negatives, these are error-repair observations rather than a benchmark-wide ranking. We omit the shared-error case breakdown here to avoid duplicating the intervention summary.

4.4.2 Known-Law Intervention

We next ask whether a correct law name is enough to repair the next stage of the judgment chain. For every model-specific baseline error on a gold-unlawful single-law image, we retain only cases where the model had correctly judged the image unlawful but named neither the gold law nor any gold provision. We then append the correct law name, while withholding the article, paragraph, and item. This yields 693 model–image targets across the 11 evaluated models.

Table 6 reports the number of targets, valid affirmative answers after the cue, and provision recovery. A hit requires the gold article and every annotated paragraph or item. Because the law name is provided in the prompt, provision matching is evaluated within that supplied law; the table therefore tests recovery of the provision rather than repetition of the cue. Technical failures and non-affirmative or malformed answers remain in the target denominator.

Table 6: Providing the correct law name, but no provision, on each model's baseline legal-grounding errors.

Model	Targets	Valid "Yes"	$\geq$1 correct provision	All correct provisions
Qwen3-VL-Plus	77	77/77	23/77 (29.9%)	15/77 (19.5%)
Qwen3-VL-Flash	72	70/72	12/72 (16.7%)	9/72 (12.5%)
DeepSeek-V4.1-Flash	37	27/37	6/37 (16.2%)	5/37 (13.5%)
GPT-4o	68	68/68	1/68 (1.5%)	1/68 (1.5%)
GPT-4o-mini	73	73/73	5/73 (6.8%)	2/73 (2.7%)
Gemini-2.5-Pro	65	64/65	18/65 (27.7%)	12/65 (18.5%)
Gemini-2.5-Flash	43	42/43	8/43 (18.6%)	4/43 (9.3%)
Grok-4.6	72	70/72	18/72 (25.0%)	9/72 (12.5%)
GLM-4.6V	68	63/68	8/68 (11.8%)	3/68 (4.4%)

Claude responses were unavailable because its endpoint remained saturated.

The law cue does not reliably recover the concrete legal basis. The largest observed all-provision counts are 15/77 (Qwen3-VL-Plus) and 12/65 (Gemini-2.5-Pro); GPT-4o and GPT-4o-mini recover all gold provisions in only 1/68 and 2/73 targets. Thus, supplying the applicable statute can repair the law-level mistake without resolving the harder article-, paragraph-, and item-level grounding problem. The rows are descriptive rather than a strict model ranking because each model contributes its own baseline-error set; Claude rows are unavailable due to endpoint saturation.

5 Conclusion

This study asked whether multimodal models can move from flagging potentially unlawful advertisements to providing defensible legal grounds. RegulaBench separates violation detection, applicable-law identification, provision grounding, and fidelity to enforcement-derived references. Across 600 images, several models detect most image-evident violations, but high recall often coincides with low specificity and a substantial false-positive burden. High sensitivity therefore does not establish reliable independent regulatory judgment.

The gap widens after the binary judgment. Exact law-name matching declines sharply for multi-law cases, even when a model identifies one relevant statute. Provision grounding is weaker still: the highest observed all-hit rate is 39.8% for single-provision images and falls to 5.5% and 2.1% for two- and three-provision images, respectively. Errors are not distributed uniformly across the legal framework. Most identifiable wrong article citations invoke the *Advertising Law*, and four familiar articles account for 63.2% of them. Principle-based grounds appearing in the case-derived references are almost absent from model errors. Outcome agreement can thus conceal incomplete coverage and reliance on a narrow set of provisions. These patterns do not, however, establish that the enforcement record is uniquely correct or that norm type itself causes citation error.

The interventions locate difficulty at multiple stages. Adding an audited description of visible scene content repairs some image-only false negatives, but the gains vary by model and are accompanied by occasional regressions or invalid outputs. Supplying the correct law name also fails to produce dependable article-, paragraph-, and item-level recovery. The remaining errors therefore cannot be explained solely by missing visual text or by failure to retrieve the statute's name; they also reflect unresolved problems in linking observed content to the appropriate legal provision.

These findings support a bounded role for multimodal models as advertising pre-regulators. They may expand screening capacity, prioritize material for human review, and organize candidate grounds, but their citations require verification and cannot replace professional judgment. The appropriate design target is therefore calibrated triage with traceable evidence, verified citations, and an explicit human decision point. The evidence is limited to generated images, image-evident labels, case-derived references, and one domain; models and human regulators did not decide identical records. Future work should test authentic advertisements and richer records, combine retrieval with provision-level verification and calibrated abstention, and test whether assistance reduces review effort without passing false positives or incomplete legal explanations downstream. Reliable AI-assisted regulation depends on preserving the chain from visual evidence to legal justification.

AI USE STATEMENT

Generative AI tools were used to construct benchmark advertising images, support research ideation, and polish the language and presentation of the manuscript. All research questions, experimental designs, analyses, interpretations, and final written content were reviewed and approved by the authors, who take full responsibility for the paper.

ETHICS STATEMENT

This study evaluates multimodal-model decisions on advertising materials derived from real administrative enforcement cases. Human administrative decisions are used as empirical institutional references rather than as guarantees of legal correctness, neutrality, fairness, or exhaustiveness. The benchmark is intended for research and evaluation of AI-assisted advertising regulation, not for determining the legal rights or liabilities of any individual or organization, certifying the legality of any advertisement, or replacing professional legal judgment. Because the study relies on real administrative materials and uses generated images that may reproduce or transform claims found unlawful in the source cases, any released data or artifacts should distinguish generated images from authentic advertisements and comply with applicable source-access, privacy, licensing, and data-use requirements.

REPRODUCIBILITY STATEMENT

The paper specifies the experimental conditions, metric definitions, case-to-image reconstruction procedure, and case-resampling approach. The appendix summarizes output checks, parsing and normalization rules, and the adjudication workflow. Any released data and artifacts will include, where legally and technically permissible, image and source identifiers, annotation and label rules, prompt and parser versions, model identifiers and access routes, evaluation scripts, and per-image outputs. Releases remain subject to applicable legal, privacy, licensing, and source-access constraints.

REFERENCES

Harethah Abu Shairah, Somayah AlHarbi, Abdulaziz AlHussein, Sameer Alsabea, Omar Shaqaqi, Hebah AlShamlan, Omar Knio, and George Turkiyyah. ALARB: An arabic legal argument reasoning benchmark. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pp. 389–406, 2025. doi: 10.18653/v1/2025.arabicnlp-main.32. URL https://aclanthology.org/2025.arabicnlp-main.32/.

Ziang Chen, Guannan Li, Fanlin Ji, Yipeng Kang, Jiaqi Li, Muhan Zhang, Yangtao Zhang, Li Tianjiao, Jiannan Wang, Xin Guo, Song-Chun Zhu, and Bin Ling. JurisBench: A deep benchmark for assessing large language models in professional legal practice. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35994–36018, 2026. doi: 10.18653/v1/2026.acl-long.1666. URL https://aclanthology.org/2026.acl-long.1666/.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. LAiW: A chinese legal large language models benchmark. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10738–10766, 2025. URL https://aclanthology.org/2025.coling-main.716/.

Ronald M. Dworkin. The model of rules. *University of Chicago Law Review*, 35(1):14–46, 1967. URL https://chicagounbound.uchicago.edu/uclrev/vol35/iss1/3/.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7933–7962, 2024. doi: 10.18653/v1/2024.emnlp-main.452. URL https://aclanthology.org/2024.emnlp-main.452/.

Rahul Ghosh, Baishali Chaudhury, Hari Prasanna Das, Meghana Ashok, Ryan Razkenari, Long Chen, Sungmin Hong, and Chun-Hao Liu. CompAgent: An agentic framework for visual compliance verification, 2026. URL https://arxiv.org/abs/2511.00171.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1915. URL https://proceedings.nips.cc/paper_files/paper/2023/hash/89e44582fd28ddfea1ea4dcb0ebbf4b0-Abstract-Datasets_and_Benchmarks.html.

Haoyang Gui, Thales Bertaglia, Taylor Annabell, Catalina Goanta, Tjomme Dooper, and Gerasimos Spanakis. Evaluating LLM-generated legal explanations for regulatory compliance in social media influencer marketing. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pp. 157–171, 2025. doi: 10.18653/v1/2025.nllp-1.12. URL https://aclanthology.org/2025.nllp-1.12/.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1705–1715, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Hussain_Automatic_Understanding_of_CVPR_2017_paper.html.

Deyi Ji, Yuekui Yang, Liqun Liu, Peng Shu, Haiyang Wu, Shaogang Tang, Xudong Chen, Shaoping Ma, Tianrun Chen, and Lanyun Zhu. RAVEN++: Pinpointing fine-grained violations in advertisement videos with active reinforcement reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1–10, 2025a. doi: 10.18653/v1/2025.emnlp-industry.1. URL https://aclanthology.org/2025.emnlp-industry.1/.

Deyi Ji, Yuekui Yang, Haiyang Wu, Shaoping Ma, Tianrun Chen, and Lanyun Zhu. RAVEN: Robust advertisement video violation temporal grounding via reinforcement reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 22–31, 2025b. doi: 10.18653/v1/2025.acl-industry.3. URL https://aclanthology.org/2025.acl-industry.3/.

Kaiyuan Ji, Yixuan Gao, Lu Sun, Yushuo Zheng, Zijian Chen, Jianbo Zhang, Xiangyang Zhu, Yuan Tian, Zicheng Zhang, and Guangtao Zhai. A^3: Towards advertising aesthetic assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9478–9490, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ji_A3_Towards_Advertising_Aesthetic_Assessment_CVPR_2026_paper.html.

Zhiwei Jia, Pradyumna Narayana, Arjun Akula, Garima Pruthi, Hao Su, Sugato Basu, and Varun Jampani. KAFA: Rethinking image ad understanding with knowledge-augmented feature adaptation of vision-language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 772–785, 2023. doi: 10.18653/v1/2023.acl-industry.74. URL https://aclanthology.org/2023.acl-industry.74/.

Kanika Kalra, Bhargav Kurma, Silpa Vadakkeeveetil Sreelatha, Manasi Patwardhan, and Shirish Karande. Understanding advertisements with BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7542–7547, 2020. doi: 10.18653/v1/2020.acl-main.674. URL https://aclanthology.org/2020.acl-main.674/.

Louis Kaplow. Rules versus standards: An economic analysis. *Duke Law Journal*, 42(3):557–629, 1992. URL https://scholarship.law.duke.edu/dlj/vol42/iss3/2/.

Renana Keydar, Noa Mor, Yuval Shany, and Omri Abend. Bending the rules: On large language models and content moderation. *Israel Law Review*, 59(1):96–133, 2026. doi: 10.1017/S0021223726100156.

Adi Levi, Or Levi, Sardhendu Mishra, and Jonathan Morra. AI vs. human moderators: A comparative evaluation of multimodal LLMs in content moderation for brand safety. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 6024–6032, 2025. URL https://openaccess.thecvf.com/content/ICCV2025W/CVAM/html/Levi_AI_vs._Human_Moderators_A_Comparative_Evaluation_of_Multimodal_LLMs_ICCVW_2025_paper.html.

Sina Malakouti, Aysan Aghazadeh, Ashmit Khandelwal, and Adriana Kovashka. Benchmarking VLMs' reasoning about persuasive atypical images. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 4788–4798, 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Malakouti_Benchmarking_VLMs_Reasoning_About_Persuasive_Atypical_Images_WACV_2025_paper.html.

Andrei Vlad Man, Răzvan-Alexandru Smădu, Cristian-George Craciun, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. RoD-TAL: A benchmark for answering questions in romanian driving license exams. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 5562–5602, 2026. doi: 10.18653/v1/2026.findings-eacl.295. URL https://aclanthology.org/2026.findings-eacl.295/.

Bill Marino, Rosco Hunter, Christoph Schnabl, Zubair Jamali, Marinos Emmanouil Kalpakos, Mudra Kashyap, Isaiah Hinton, Alexa Hanson, Maahum Nazir, Felix Steffek, Hongkai Wen, and Nicholas D. Lane. AIReg-Bench: Benchmarking language models that assess ai regulation compliance, 2025. URL https://arxiv.org/abs/2510.01474.

Pouria Rad, Ahmed Aleroud, Gianluca Zanella, and Lin Li. AI-powered decision support in privacy laws: Benchmarking LLMs on legal violation detection. In *Proceedings of the 59th Hawaii International Conference on System Sciences*, 2026. doi: 10.24251/HICSS.2026.819. URL https://doi.org/10.24251/HICSS.2026.819.

Diya Saha, Abhishek Bharadwaj Varanasi, Tirthankar Dasgupta, and Manjira Sinha. RegNLI: Detecting online product misbranding through regulatory and linguistic alignment. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 417–424, 2026. doi: 10.18653/v1/2026.eacl-industry.31. URL https://aclanthology.org/2026.eacl-industry.31/.

Makoto Takamatsu, Modong Tan, Ryotaro Sugiura, Mioko Hirose, and Naoya Takizawa. Towards explainable AI in advertising compliance: A logic-driven two-stage multi-agent system. Extended abstract, AI Governance: Alignment, Morality and Law Workshop at AAAI 2025, 2025. URL https://openreview.net/pdf/334a39852367bccaef22875f074c178e94fc0832.pdf.

Shreyash Verma, Amit Kesari, Vinayak Trivedi, Anupam Purwar, and Ratnesh Jamidar. M-PACE: Mother child framework for multimodal compliance, 2025. URL https://arxiv.org/abs/2509.15241.

Yiran Yang, Zhaowei Liu, Yuan Yuan, Yukun Song, Xiong Ma, Yinghao Song, Xiangji Zeng, Lu Sun, Yulu Wang, Hai Zhou, Shuai Cui, Zhaohan Gong, and Jiefei Zhang. BLM-Guard: Explainable multimodal ad moderation with chain-of-thought and policy-aligned rewards. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35985–35993, 2026. doi: 10.1609/aaai.v40i42.40914. URL https://ojs.aaai.org/index.php/AAAI/article/view/40914.

Keren Ye and Adriana Kovashka. ADVISE: Symbolism and external knowledge for decoding advertisements. In *Proceedings of the European Conference on Computer Vision*, 2018. URL https://www.openaccess.thecvf.com/content_ECCV_2018/papers/Keren_Ye_ADVISE_Symbolism_and_ECCV_2018_paper.pdf.

A Additional Results

A.1 Detailed Violation-Judgment Results

Table 7 provides the response-coverage, output-validity, count, and uncertainty details underlying the summary results in Table 1.

Table 7: Image-level violation-judgment performance. For complete models, accuracy uses all 600 images, recall the 400 image-evident violation cases, and specificity the 200 human-reviewed images with no image-evident violation. For † provisional rows, each metric uses only API-successful images in its corresponding stratum. Accuracy entries are correct/N (% [95% Wilson CI]); recall and specificity entries are % [95% Wilson CI] (correct/explicit error).

Family	Model	API	Invalid[a]	Accuracy	Recall (TP/FN)	Specificity (TN/FP)
Qwen	Qwen3-VL-Plus	600/600	0 (0/0)	472/600 (78.7 [75.2, 81.8])	99.8 [98.6, 100.0] (399/1)	36.5 [30.1, 43.4] (73/127)
Qwen	Qwen3-VL-Flash	600/600	2 (2/0)	493/600 (82.2 [78.9, 85.0])	96.2 [93.9, 97.7] (385/13)	54.0 [47.1, 60.8] (108/92)
DeepSeek	DeepSeek-V4.1-Flash	600/600	21 (4/17)	447/600 (74.5 [70.9, 77.8])	98.0 [96.1, 99.0] (392/4)	27.5 [21.8, 34.1] (55/128)
OpenAI	GPT-4o	600/600	0 (0/0)	528/600 (88.0 [85.2, 90.4])	90.5 [87.2, 93.0] (362/38)	83.0 [77.2, 87.6] (166/34)
OpenAI	GPT-4o-mini	600/600	0 (0/0)	473/600 (78.8 [75.4, 81.9])	95.2 [92.7, 96.9] (381/19)	46.0 [39.2, 52.9] (92/108)
Google	Gemini-2.5-Pro	600/600	0 (0/0)	423/600 (70.5 [66.7, 74.0])	100.0 [99.0, 100.0] (400/0)	11.5 [7.8, 16.7] (23/177)
Google	Gemini-2.5-Flash	600/600	15 (15/0)	448/600 (74.7 [71.0, 78.0])	96.0 [93.6, 97.5] (384/1)	32.0 [25.9, 38.8] (64/136)
Anthropic	Claude-Sonnet-4.5†	584/600	37 (27/10)	454/584 (77.7 [74.2, 80.9])	89.0 [85.6, 91.7] (356/17)	53.3 [46.1, 60.3] (98/76)
Anthropic	Claude-Haiku-4.5†	584/600	129 (70/59)	372/584 (63.7 [59.7, 67.5])	81.5 [77.4, 85.0] (326/4)	25.0 [19.3, 31.7] (46/79)
xAI	Grok-4.6	600/600	0 (0/0)	508/600 (84.7 [81.6, 87.3])	98.8 [97.1, 99.5] (395/5)	56.5 [49.6, 63.2] (113/87)
GLM	GLM-4.6V	600/600	4 (2/2)	495/600 (82.5 [79.3, 85.3])	94.0 [91.2, 95.9] (376/22)	59.5 [52.6, 66.1] (119/79)

[a] API-successful outputs without an exact, machine-readable binary judgment; parentheses give invalid counts on violation/non-violation images. Invalid judgments count as incorrect in rate denominators. TP/FN and TN/FP count valid binary judgments only, so the corresponding invalid count completes each stratum denominator. API failures are coverage failures rather than model errors. †Provisional rows use current API-successful outputs and are excluded from the main-table ranking.