ScreenshotLegalBench: A Multimodal Benchmark for Legal Evidence Understanding in Chat Screenshots

Anonymous EMNLP submission

Abstract

Chat screenshots from platforms such as WeChat are increasingly used as legal evidence in Chinese civil litigation. However, their informal layout, multimodal nature, and lack of structure pose significant challenges for automated understanding. We introduce **ScreenshotLegalBench**, the first large-scale multimodal benchmark for *Legal Screenshot Evidence Understanding (LSEU)*. It supports two key tasks: (1) structured key information extraction (KIE) and (2) legal visual question answering (VQA). The dataset contains over 4,600 chat screenshots annotated with 145,044 structured labels, a 143-image evaluation set with 2,678 verified annotations, and 1,176 VQA instances covering evidence relevance, format validity, and legal reasoning. Among these, 106 cases involve multi-image cause-of-action scenarios. We benchmark several open-source vision-language models (VLMs), including InternVL and Qwen-VL families. Experimental results show that current VLMs struggle with layout interpretation and domain-specific reasoning, despite instruction tuning. **ScreenshotLegalBench** offers a novel and scalable resource at the intersection of vision, language, and law, enabling future research on multimodal legal document understanding in real-world settings. The dataset and code are soon publicly available at Github.

1 Introduction

In recent years, multimodal large models have emerged as a major focus in AI research, demonstrating impressive performance on tasks that require integrating image and text modalities.Within the legal domain, such models are increasingly expected to automate the preprocessing of complex and loosely structured case materials, particularly in civil proceedings where parties are required to submit supporting evidence in diverse digital formats. Among these, WeChat chat screenshots have become a common form of statutory evidence in China, as officially recognized in Art.116 of the Supreme People's Court's Interpretation on the Civil Procedure Law.

However, analyzing chat screenshots remains a deeply manual task in judicial practice. Legal practitioners must verify speaker identities, reconstruct conversation sequences, determine the legal relevance of each message, and evaluate whether the screenshot satisfies evidentiary requirements. Unlike formal documents, chat screenshots are informal, heterogeneous, and visually irregular—mixing text, images, emojis, file attachments, and transfer records within complex UI layouts. The absence of structured representations, coupled with the multimodal and noisy nature of the content, poses significant obstacles to automation. As a result, lawyers and judges must perform labor-intensive and error-prone manual reviews, which reduces efficiency and increases the risk of oversight.

To address this gap, we propose the task of *Legal Screenshot Evidence Understanding (LSEU)*, which aims to extract structured legal information and assess evidentiary attributes from real-world, multimodal inputs. Unlike generic visual question answering (VQA) or document understanding tasks, LSEU presents unique challenges: (1) informal and irregular visual layouts, (2) field-level legal structuring grounded in procedural norms, and (3) domain-specific reasoning required for admissibility assessment and case interpretation. As illustrated in Appendix Figure 3, legal practitioners typically process such evidence in five stages: relevance screening, timeline reconstruction, factual extraction, legal mapping, and litigation strategy development. Our benchmark focuses on three of these stages, namely structured perception, factual extraction, and legal mapping, which collectively capture the core components of real-world judicial workflows.

However, existing vision-language models struggle to address LSEU effectively due to three key challenges: (1) the multi-modal and informal nature of chat screenshots, (2) the need for structured field extraction aligned with legal interpretation, and (3) the domain-specific reasoning required to infer case types or assess evidentiary admissibility.

To tackle these challenges, we decompose the LSEU task into two core components: **structured perception**, which detects and extracts legally relevant information (e.g., who said what, when, and with what legal implication), and **semantic reasoning**, which classifies legal relevance, determines evidentiary status, and infers preliminary case hypotheses. This decomposition reflects how legal professionals process chat evidence—first segmenting and organizing visual information, then reasoning over the structured content. **We are guided by the following research questions:**

- *RQ1: Can large vision-language models accurately extract legally structured information from noisy, multimodal chat screenshots?*
- *RQ2: Does a two-stage modeling pipeline that combines structured perception with downstream semantic reasoning outperform end-to-end baselines in legal classification and case reasoning tasks?*

To answer these questions, we present **ScreenshotLegalBench**, the first publicly available benchmark designed for multimodal legal evidence understanding in chat-based scenarios. The dataset includes over 4,600 WeChat chat screenshots annotated for KIE, as well as more than 1,100 VQA pairs targeting legal classification, evidence validity, and cause-of-action reasoning. All annotations are performed by certified Chinese legal professionals, ensuring alignment with practical legal standards.**Our contributions are summarized as follows:**

- We introduce **ScreenshotLegalBench**, a multimodal benchmark for legal evidence understanding in chat screenshots, comprising real-world and simulated samples annotated by certified Chinese legal professionals to reflect practical legal needs and legal reasoning demands.
- We introduce **Legal Screenshot Evidence Understanding (LSEU)** as a two-stage task: structured KIE and legal VQA, supported by a unified annotation schema and scalable labeling pipeline.
- We benchmark a range of open-source vision-language models under realistic deployment constraints, focusing on models that are suitable for local and privacy-sensitive legal environments. The results show that even advanced instruction-tuned models continue to face persistent challenges in layout robustness, field-level structuring, and legal reasoning across diverse input forms.

By bridging the domains of vision, language, and legal reasoning, this work offers a new foundation for multimodal legal AI. We hope that **ScreenshotLegalBench** will catalyze future research on interpretable and reliable systems for real-world evidence analysis.

2 Related Work

Legal Information Extraction. Early work in legal NLP focused on clause extraction and entity identification from structured contracts or rulings. CUAD (Hendrycks et al., 2021), LEDGAR (Tuggener et al., 2020), and ACORD (Wang et al., 2025) provide high-quality text-based datasets for commercial clause classification and retrieval. However, these datasets operate on well-formatted, language-only documents, lacking support for multimodal input or visual layout reasoning.

Recent studies(Liu et al., 2023) show that incorporating visual cues such as bounding boxes and font styles can improve structured extraction from long documents. These findings motivate our structured KIE approach for visually noisy chat screenshots, which lack standardized layouts and often contain overlapping modalities such as image messages, emojis, and file transfers. This setting poses new challenges for entity alignment and layout-robust modeling.

Legal Reasoning and Understanding. Benchmarks like LexGLUE (Chalkidis et al., 2022) and CaseHOLD (Zheng et al., 2021) define a suite of judgment prediction, retrieval, and question answering tasks. LEGAL-BERT (Chalkidis et al., 2020) demonstrates the importance of domain-adaptive pretraining. However, these benchmarks assume clean, pre-extracted legal facts and do not support evidence-level interpretation from raw multimodal inputs.

Chinese datasets such as JEC-QA(Zhong et al., 2019) and CAIL2018(Xiao et al., 2018) provide useful resources for statutory reasoning and charge prediction but remain focused on formal court doc-

Task Type	Data Concern	Structured Layout	Visual Reasoning	Textual Semantics	Multimodal Alignment	Temporal Reasoning	Legal Action Modeling
Form KIE	Doc Image + OCR + Position	✓	✗	△	✗	✗	✗
Layout Parsing	Doc Image + Layout Tags	✓	✓	△	✓	✗	✗
DocVQA	Doc Image + OCR + QA Pair	△	△	✓	△	✗	✗
TextVQA	Scene Image + OCR + Question	✗	✓	✓	△	✗	✗
Table QA	Table Image + Question	✓	✗	✓	△	✗	✗
InfoVQA	Image + OCR + Embedded Info	△	✓	✓	✓	✗	✗
LSEU (Ours)	Screenshot + OCR + Bubble + Media + Timestamps	△	✓	✓	✓	✓	✓

Table 1: Comparison of our chat screenshot task with representative KIE/VQA tasks. ✓ indicates presence of the feature; ✗ indicates absence; △ indicates partial support or context-dependent presence.

uments. In contrast, our task addresses an earlier stage of legal workflows—assessing the admissibility and relevance of raw WeChat evidence prior to fact consolidation. This pre-factual focus introduces challenges in determining whether a screenshot constitutes legal evidence at all, requiring both structural perception and contextual inference.

Recent systems such as MASER (Jeon et al., 2022) demonstrate the ability of MLLMs to infer event chronology under weak timestamp supervision. While such systems highlight the potential of multimodal legal reasoning, they typically operate on clean, structured records, and do not handle the fragmented, layout-rich format of chat screenshots.

Multimodal Legal Benchmarks. To date, few datasets address multimodal legal evidence analysis. Prior work in DocVQA (Mathew et al., 2021), TextVQA (Hegde et al., 2023), and InfoVQA (Mathew et al., 2022) has explored visual-text reasoning in document or scene settings, but these benchmarks lack legal-specific labels and structural alignment requirements. In contrast, our task focuses on raw WeChat chat screenshots, which are often unstructured, multimodal, and legally ambiguous. It integrates both structured KIE and VQA, covering interface layout, temporal ordering, and high-level legal implications. Table 1 provides a comparative summary of our task relative to representative KIE/VQA benchmarks across multiple reasoning dimensions.

Our proposed dataset, ScreenshotLegalBench, is the first to support three interrelated subtasks over real-world chat screenshots: (1) *structured KIE* (speaker, content, time, etc.); (2) *legal attribute classification* (e.g., relevance and evidentiary status); and (3) *open-ended cause-of-action generation*. These capabilities reflect practical needs in legal workflows such as fact triage, evidence validation, and dispute summarization, which are tasks that precede traditional legal judgment prediction.

3 Task Definition

We define two core tasks for **Legal Screenshot Evidence Understanding (LSEU)**, reflecting the structured perception and legal reasoning stages over chat-based visual evidence.

Task 1: *Screenshots Evidence Key Information Extraction (SEKIE)* aims to extract structured legal fields from a single WeChat chat screenshot. The model must jointly understand the layout and semantics of the chat interface and output a structured JSON record containing message-level fields. The expected fields include:

- *speaker*: the display name of the message sender;
- *timestamp*: the message time, if available;
- *content*: the textual content of the message;
- *message_bbox*: the bounding box of the message region;
- *transfer*, *image*, *file*: optional fields indicating the presence and description of funds transfer, images, or file attachments.

This task forms the structural foundation for downstream classification and reasoning.

Task 2: *Chat Screenshot Legal Visual Legal Question Answering (CSLVQA)* evaluates the model's ability to perform higher-level legal understanding based on the screenshot. It includes three subtasks: (1) classifying whether the screenshot is legally relevant (`classify`); (2) judging whether the screenshot qualifies as formally valid evidence (`evidence`); and (3) generating a natural language description of the underlying dispute or legal issue (case_text). Notably, this sub-task is a multi-image reasoning task, where the model must synthesize information across multiple screenshots to infer a coherent legal cause.The VQA task can be performed either directly from the raw image or using the structured output from KIE as additional input.

4 The ScreenshotLegalBench Dataset

We construct a dataset for Legal Evidence Understanding in Chat Screenshots.

4.1 Data Collection

We initially obtain raw images over 9,800 candidate images from the Gansu Provincial Digital Rule of Law Industry Research Academy ,which are from Common Crawl, Google and Baidu search results, and internal institutional repositories. Each image, along with its embedded caption, was processed by Gemini 1.5 (Team et al., 2024) to determine whether it resembled a chat interface and to assign a coarse-grained content label (e.g.,"startup," "divorce," "romantic relationship"). This automated step yielded approximately 6,000 images likely to depict chat screenshots. Subsequently, trained annotators manually reviewed the topic labels and visual layout to identify cases with potential legal relevance, resulting in a filtered set of 4,800s samples for annotation in ScreenshotLegalBench.

4.2 Annotation Data Elicitation

We elicited the ScreenshotLegalBench annotations by defining two complementary pipelines for our KIE and VQA tasks (see Appendix B for full schema and annotation guidelines).Figure 4 illustrates a sample of the task annotations.

KIE task employs YOLOv3 (Redmon and Farhadi, 2018), DETR (Carion et al., 2020), and Cascade R-CNN (Cai and Vasconcelos, 2017), fine-tuned on a 50-image few-shot subset, to localize 16 key interface elements, including message bubbles, avatars, timestamps, and transaction indicators. Targeted data augmentation enhances robustness across diverse chat layouts, yielding a 70.1% mAP. These detectors identify candidate layout regions across the dataset, from which text is extracted using PaddleOCR and Google OCR. Multi-line transaction entries are semantically merged through rule-based consolidation into coherent legal statements (e.g., a three-line WeChat transfer becomes "WeChat transfer received ¥520.00, note: happy birthday"). File-related content is normalized by parsing filenames and extensions, while image-only regions are annotated using a multimodal generative model to enhance reasoning context.Speaker attribution is determined by analyzing bounding-box centroids relative to the page's vertical axis, assigning right-side elements to the primary speaker and left-side elements to the interlocutor. Missing timestamps are interpolated using a sliding-window strategy to maintain temporal continuity. All spatial, textual, and semantic annotations are organized into a unified JSON schema, serving as high-quality weak supervision for downstream causal analysis and evidentiary chain reconstruction. The layout schema and bounding-box definitions are detailed in Appendix B.

For the **VQA task**, we designed a set of expert-authored legal questions to elicit complex reasoning abilities from multimodal models. **Notably**, all VQA annotations were created from scratch by experienced legal professionals, without the use of automated pre-labeling. The questions are divided into two levels: global questions, which assess the screenshot as a whole, and local questions, which focus on fine-grained content such as individual messages or UI elements. The **global questions**, which have been fully annotated and released, guide the model to reason across four legal dimensions: (a) whether the image is a chat screenshot and holds legal relevance; (b) whether it satisfies evidentiary completeness and admissibility criteria; (c) whether it depicts private or group conversation; and (d) what type of legal dispute (e.g., loan, labor) the conversation may suggest. These questions serve as the foundation for high-level evidence screening and case framing. In contrast, the **local questions** target specific visual or textual components such as message content, avatars, quoted speech, emojis, transfers, and file attachments. They are designed to test the model's ability to extract intent, recognize legal relationships, classify transaction types, and interpret symbolic or emotive cues. Due to their labor-intensive nature, local annotations are still in progress and will be released in a future update alongside detailed statistics. Nonetheless, the schema has been finalized to ensure backward compatibility and extensibility. Table 6 presents representative examples of local-level questions and their annotation goals.

4.3 Manual Correction and Expert Review

To ensure evaluation integrity, we conducted detailed manual correction for a held-out subset of 143 KIE proposals. These proposals were initially generated by baseline models and subsequently reviewed by trained annotators. Corrections included bounding box adjustments for spatial accuracy, merging or splitting of entity spans, and fixing misclassified field types. This process yielded a reliable reference set for evaluating KIE performance. Full annotation criteria and workflows are provided in Appendix C.

For the VQA task, all annotations were fully manual. Expert annotators first created bounding boxes and question–answer pairs based on predefined prompts.All annotations were performed by legally qualified annotators who had passed China's National Judicial Examination. For questions requiring nuanced legal judgment, responses were validated by senior attorneys with over a decade of practice, ensuring consistency with real-world legal reasoning.

4.4 Annotation Quality Assurance

To ensure the consistency, completeness, and legal validity of ScreenshotLegalBench annotations, we implemented a multi-stage quality assurance protocol integrating model-assisted pre-processing, a hierarchical annotation framework, and multi-level expert review.

As described in Sections 4.2 and 4.3, both the KIE and VQA pipelines combine structured interfaces, heuristic post-processing, and expert-in-the-loop validation. For KIE, model-generated layout elements were aligned with OCR results and then refined through legal-specific consolidation and manual correction on 143 samples. For VQA, all annotations were manually created, with legally sensitive questions reviewed by senior attorneys.

As shown in Table 5, this framework begins with global property tagging (e.g., legality, chat type, case type) and progresses to layout-level detection (e.g., message, avatar, file), content structuring (e.g., speaker name, message text), semantic transformation (e.g., merging transfer info into coherent legal phrases), and finally to legal question answering over both global and local visual regions.Each level corresponds to a distinct layer of information abstraction required for multimodal legal understanding.

This layered structure ensures both fine-grained supervision for information extraction and high-level signals for reasoning tasks. All annotations followed centralized task formats, and ambiguous cases were discussed and resolved through collaborative expert review. Examples of annotation interfaces and VQA samples are provided in Appendix B.5.

4.5 Dataset Statistics

ScreenshotLegalBench comprises three complementary subsets designed to support multimodal legal tasks in WeChat chat screenshots: (1) an object detection subset for layout element localization, (2) a large-scale KIE corpus for pretraining and evaluation, and (3) a VQA benchmark for multimodal legal reasoning. Notably, the KIE and most VQA tasks are annotated at the single-image level, while the `case_text` sub-task adopts a multi-image setting—each case aggregates an average of 4.7 screenshots to support cause-of-action analysis across dialogue contexts. Overall statistics are summarized in Table 2, with detailed field counts provided in Appendix C.

5 Experiments

We evaluate a range of vision–language models on ScreenshotLegalBench to assess their performance on structured perception and legal reasoning. Section 5.3 compares open-source baselines on KIE and VQA tasks. Section 5.4 demonstrates that, even with partially automated annotations, fine-tuned models outperform larger zero-shot baselines, highlighting the generalizability of our dataset. Finally, in Section 5.5, we conduct ablations to validate our dataset design, demonstrating that the inclusion of KIE as a structured perception task significantly improves downstream legal classification and reasoning.

5.1 Benchmark Models

Early legal NLP systems often relied on rule-based heuristics or traditional machine learning (e.g. SVMs(Chen and Lin, 2006)), but these methods tend to fail under the noisy, layout-rich conditions of chat screenshots. Our evaluation focuses on multimodal foundation models with strong image–text processing capabilities, excluding shallow baselines. All experiments are conducted under data confidentiality constraints, dueing to real-world deployment needs in local, privacy-sensitive legal environments.We select two prominent model fam-

Subset	Samples	Total Annotations	Avg. Annotations per Sample	Main Tasks
Object Detection	50 screenshots	945 bounding boxes	18.9	Layout Element Detection
KIE Training Set	39,477 messages	145,044 fields	3.67	Structured Pretraining
KIE Eval Set	143 screenshots / 696 messages	2,678 fields	3.85	Structured Evaluation
VQA Set	1,176 screenshots / 106 multi-imgs case	2,854 legal annotations	2.42 legal QA pairs	Dialog Type, Evidence, Case Reasoning

Table 2: Overview of ScreenshotLegalBench dataset subsets.

ilies,InternVL (Chen et al., 2024c,b,a) and Qwen-VL (Bai et al., 2023a; Wang et al., 2024; Bai et al., 2025), as benchmark decoders. Both are widely used in Chinese image–text tasks and are capable of generating structured outputs:

- **InternVL2** (Chen et al., 2024b)is a dual-encoder model fine-tuned on ScreenshotLegalBench for legal information extraction.
- **InternVL2.5**(Chen et al., 2024a) extends InternVL2 with QLoRA for efficient domain adaptation, updating only low-rank adapters while freezing the visual and language backbones.
- **Qwen2-VL**(Wang et al., 2024) and **Qwen2.5-VL**(Bai et al., 2025) are Transformer-based models optimized for Chinese image–text inputs, pretrained on multilingual corpora and capable of structured output generation.

All models are evaluated using the prompt templates detailed in Appendix D.4. Fine-tuned models are assessed under `pass@0` to reflect output stability, while raw models are evaluated under `pass@1`. Higher format scores indicate stronger structural adherence and benefit from evaluation-side repair strategies enhancing JSON compatibility.

5.2 Evaluation Metrics

We evaluate model performance separately on the KIE and VQA tasks. For KIE, the output is a structured JSON containing a list of messages, each with textual and spatial fields. We measure quality along three axes: (1) structural validity, checking that each message includes all required fields in legal formats; (2) semantic accuracy, computed via a hybrid similarity score that combines normalized token-wise alignment and substring overlap; and (3) spatial alignment, evaluated using standard **Intersection-over-Union (IoU)** between predicted and reference bounding boxes. The semantic score for a message field y against reference $\hat{y}$ is defined as

$$\mathrm{Sim}(y, \hat{y}) = \lambda\, \mathrm{SeqSim}(y, \hat{y}) + (1 - \lambda)\, \mathrm{LCS}(y, \hat{y})$$

where SeqSim measures the proportion of aligned token spans under optimal matching, and LCS denotes the ratio of the longest common substring length. Overall KIE score averages across valid messages.

For VQA, we consider two classification tasks (`classify`, `evidence`) and one generation task (`case_text`). The classification tasks are evaluated using macro-averaged Precision, Recall, F1, and Accuracy, as the label distributions are notably imbalanced. For `case_text`, we evaluate the ability of the model to generate a concise, legally coherent description of the dispute based solely on the screenshot.While full-text metrics such as BLEU(Papineni et al., 2002) or ROUGE(Lin, 2004) are widely used in generative tasks, they are ill-suited for legal cause-of-action summaries due to the professional phrasing, variable expression, and high semantic abstraction involved. Instead, we adopt a simplified but interpretable metric: hit rate over legal dispute categories, which evaluates whether the predicted output contains at least one correct category keyword. Formally, the metric is defined as

$$\text{Dispute HitRate} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\exists\, c \in C_i \cap \hat{C}_i]$$

where C_i is the set of dispute keywords extracted from the model output and $\hat{C}_i$ is the gold label set. A hit is counted if at least one legal category is correctly recovered. Additional metrics such as normalized similarity and output length consistency are used for robustness and are detailed in Appendix B.

5.3 Main Performance

KIE Task Results We evaluate vision–language models on the KIE task using ScreenshotLegalBench (Figure 1, Table 10). The task evaluates structured output quality, focusing on format validity, spatial alignment (IoU), and content accuracy.

As shown in Table 10, performance varies significantly across models. InternVL2.5-2B(Chen et al., 2024a) achieves the highest overall score of 0.6302,

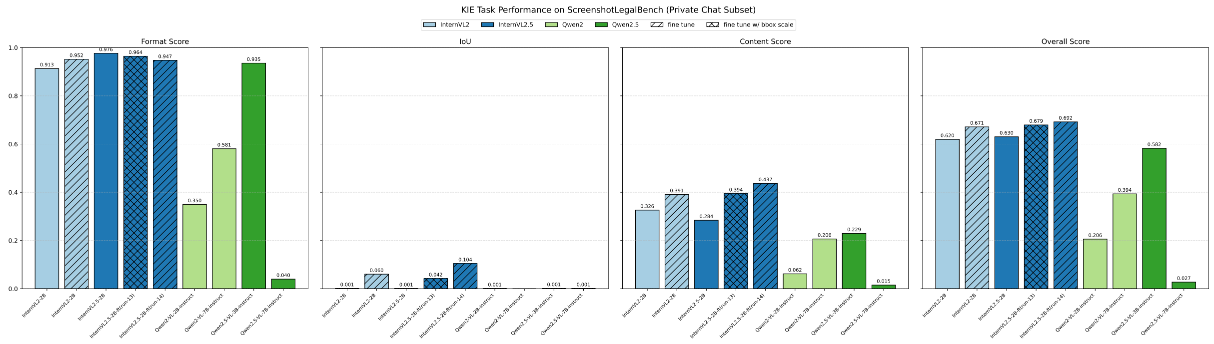


Figure 1: Comparison of KIE task performance on ScreenshotLegalBench (Private Chat Subset).Fine-tuned models are evaluated under pass@0, which better reflects output stability, while other models are evaluated under pass@1. Higher Format Scores indicate not only stronger adherence to structural output instructions, but also reflect the contribution of evaluation-side repair strategies designed to maximize compatibility with JSON-based outputs.

with strong format validity (0.9764). However, spatial alignment remains a challenge for most models, with un-tuned models, including Qwen2-VL-7B-instruct(Wang et al., 2024), showing near-zero IoU, indicating poor spatial reasoning. Content accuracy also varies, with InternVL2.5-2B(Chen et al., 2024a) scoring 0.2839, while larger models like Qwen2.5-VL-7B-instruct(Bai et al., 2025) score much lower (0.0151).

These results highlight the complexity of generating structured legal outputs and the need for fine-tuning, which is further explored in the next section on dataset generalization.

VQA Task Results Unlike the KIE task that emphasizes local timeline reconstruction under privacy constraints and is evaluated with smaller models, the VQA task targets legal reasoning and evidentiary judgment, requiring more complex abstraction. Therefore, larger-scale vision–language models are included to better assess their legal understanding capabilities. We evaluate model performance on three sub-tasks in the VQA portion of ScreenshotLegalBench: (1) legal relevance classification (`classify`), (2) assessment of evidentiary format compliance (`evidence`), and (3) cause-of-action generation via multi-image reasoning (`case_text`),which detail in 2. As shown in Table 11, classification performance is low across models (F1 < 0.05), reflecting the difficulty of determining legal relevance without contextual cues. The *evidence* task remains especially challenging: even large models like Qwen2.5-VL-72B(Bai et al., 2025) achieve high recall (0.44) but near-zero precision, indicating poor understanding of legal format standards. In *case_text* (Table 12), models fail to produce coherent multi-image legal summaries. Larger models (e.g., 72B) do not outperform smaller, instruction-tuned variants, suggesting that scale alone is insufficient for legal abstraction.

In summary, these results yield three key insights: (1) current models struggle to assess evidentiary formality due to limited spatial and layout understanding, and (2) moderate-sized models with domain-specific tuning outperform larger zero-shot models on multi-image legal reasoning tasks.

5.4 Dataset Generalization Analysis

To assess the benefits of dataset-specific instruction tuning, we compare the performance of foundation models and their fine-tuned counterparts on the KIE task (Figure 1). Fine-tuned models exhibit significantly higher content accuracy and spatial alignment scores, particularly in average IoU, confirming that domain-specific fine-tuning enhances the model's structural consistency and understanding of legal semantics.

5.5 Ablation Analysis

We conduct ablation studies to examine how different components and dataset design choices affect model performance on ScreenshotLegalBench, highlighting the benefits of structured supervision for both KIE and VQA tasks.

Importance of Bounding Box Fields (KIE). We further study the impact of providing message-level bounding boxes (`message_bbox`) during training. As shown in Table 15, removing this field leads to a near-zero IoU and degraded overall scores, despite only slight changes in content accuracy. This suggests that spatial annotations are critical for enabling the model to align textual fields with their visual locations, which is vital for downstream ap-

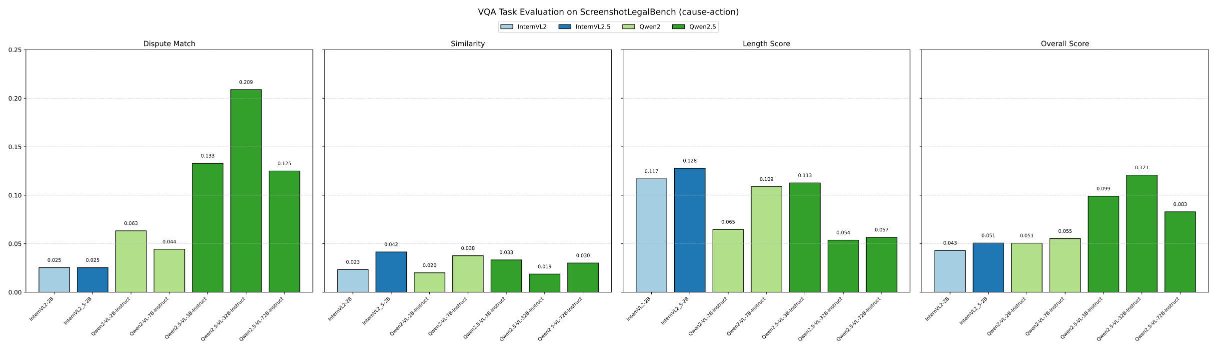


Figure 2: Comparison of vision-language models on the **case_text** VQA task in ScreenshotLegalBench.

Model	**Format Score**	**Avg. IoU**	**Content Score**	**Overall Score**
InternVL2-2B (w/ bbox)	0.9384	0.0338	0.3703	0.6544
InternVL2-2B (w/o bbox)	0.7579	0	0.3934	0.5756

Table 3: Performance of InternVL2-2B on KIE task with and without `message_box` bounding boxes

Setting	Task	Accuracy	Macro P	Macro R	F1 Score
w/o KIE guidance	classify	0.4286	0.3333	0.1428	0.1999
+KIE-augmented input	classify	**0.8333**	**0.5000**	**0.4167**	**0.4545**
w/o KIE guidance	evidence	0.0157	0.0013	0.0107	0.0019
+KIE-augmented input	evidence	**0.0187**	**0.0126**	**0.0606**	**0.0204**

Table 4: Ablation: Effect of structured KIE input on VQA tasks.

plications such as evidence localization and timeline reconstruction.

Structured vs. Plain Inputs (VQA). We assess the impact of structured perception by comparing two configurations: models predicting directly from raw screenshots (*w/o KIE guidance*) and those augmented with structured fields from the fine-tuned KIE module (*+KIE-augmented input*). As shown in Table 14, structured input significantly improves performance in the `classify` task—accuracy rises from 42.86% to 83.33%, and macro F1 more than doubles, demonstrating the value of upstream legal structuring. In contrast, the `evidence` task remains challenging. Despite slight gains from KIE augmentation, performance is low across the board. This suggests that models struggle to internalize evidentiary standards without domain-specific training, and that format validity requires not just structural cues but legal commonsense—still absent in current MLLMs. This ablation uses prompts enhanced with KIE outputs from our best-performing fine-tuned model, evaluated on Qwen-VL-MAX(Bai et al., 2023b). While limited to single-image inputs, future work should explore multi-image reasoning (e.g., `case_text`) once token constraints are addressed. Full settings and prompts are in Appendix D.4.

6 Conclusion

We present **ScreenshotLegalBench**, a new benchmark designed for legal evidence understanding in WeChat chat screenshots. It focuses on structured perception and legal reasoning tasks, offering insights into the challenges of multimodal legal AI. Despite limitations such as annotation consistency and scalability, the dataset provides a solid foundation for research in secure and practical legal applications. Baseline results reveal the difficulty of this task for current models, particularly under local deployment constraints. We encourage future work on improving scalability, layout robustness, and real-world adaptability. Positioned at the intersection of natural language processing, computer vision, and legal reasoning, LSEU holds substantial practical relevance. The dataset is released to support progress toward AI systems capable of interpreting digital legal evidence with accuracy and transparency.

Limitations

ScreenshotLegalBench presents several limitations. Although annotations are verified by legal experts,The dataset exhibits category imbalance, particularly in cause-of-action types and funds-transfer content, as most samples originate from a narrow range of civil disputes. Timestamp labels rely on visual order assumptions (based on legal experts' experience), which may be unreliable in real-world scenarios. The evaluation favors structured JSON outputs and may penalize models with strong semantics but poor formatting.

Ethics Statements

The ScreenshotLegalBench dataset is constructed using publicly available web data sourced from Gansu Provincial Digital Rule of Law Industry Research Academy . It was gathered with the intention of facilitating local, privacy-sensitive legal AI deployments, particularly for the KIE tasks. The dataset is designed to aid the automation of legal workflows while ensuring compliance with data privacy and confidentiality standards, especially in legal contexts.

To protect privacy, anonymization procedures are applied to identifiable data, such as the use of "avator_1" and "avator_0" to mask avatars in chat screenshots. These identifiers do not correspond to any real-world individual and are used solely for the purpose of maintaining privacy. However, due to the nature of web scraping, certain non-textual content in the images (e.g., emoticons, background images) and some personal information may not be entirely anonymized. Moreover, due to the structure of the raw web data, efforts to mask or obscure personal identifiers in the images (such as applying blur or cropping) may negatively affect the understanding of the primary content, as it could lead to distortion or removal of critical evidence.

The dataset's open-source release is aimed at enabling local model deployments for legal practitioners who may not have access to proprietary AI models due to regulatory or privacy concerns. This ensures that users can access advanced AI tools while retaining full control over the data and the models they develop.

While efforts have been made to ensure the privacy of the data, there may be inherent risks associated with using this dataset, especially regarding the presence of potentially noisy data, which could affect model performance. It is important for future users of the dataset to be aware of these limitations and the trade-off between privacy preservation and data quality.

As part of the ongoing ethical commitment, we also provide a mechanism for obtaining consent for the data used. Any additional requests for sensitive data or further clarifications regarding the use of this dataset can be directed to the dataset's licensing terms, with the option to obtain permissions from the data providers where necessary.

The dataset is provided solely for academic research and benchmarking purposes. Commercial use or deployment in production environments is not permitted. We hope that the ScreenshotLegalBench dataset will contribute to the development of responsible, transparent, and privacy-conscious AI systems for legal tasks, while fostering further advancements in multimodal legal document understanding.

Acknowledgments

We thank Beijing Hairun Tianrui (Zhengzhou) Law Firm for their support in annotation work. We also thank Zhangzhouliang and the Shanghai AI Lab for providing computational resources. Finally, we appreciate the anonymous reviewers for their valuable feedback.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023a. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv: 2308.12966*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv: 2502.13923*.

Zhaowei Cai and Nuno Vasconcelos. 2017. Cascade r-cnn: Delving into high quality object detection.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. *European Conference on Computer Vision*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english.

Yi-Wei Chen and Chih-Jen Lin. 2006. *Combining SVMs with Various Feature Selection Strategies*, pages 315–324. Springer Berlin Heidelberg, Berlin, Heidelberg.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv: 2412.05271*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Cvpr*.

Shamanthak Hegde, Soumya Jahagirdar, and Shankar Gangisetty. 2023. Making the v in text-vqa matter. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5580–5588.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review.

Jeewon Jeon, Woojun Kim, Whiyoung Jung, and Youngchul Sung. 2022. Maser: Multi-agent reinforcement learning with subgoals generated from experience replay buffer. *arXiv preprint arXiv: 2206.10607*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *NEURIPS*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush

Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary

Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1235–1241, Marseille, France. European Language Resources Association.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv: 2409.12191*.

Steven H Wang, Maksim Zubkov, Kexin Fan, Sarah Harrell, Yuyang Sun, Wei Chen, Andreas Plesner, and Roger Wattenhofer. 2025. Acord: An expert-annotated retrieval dataset for legal contract drafting. *arXiv preprint arXiv:2501.06582*.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv: 1807.02478*.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 159–168, New York, NY, USA. Association for Computing Machinery.

Haoxiang Zhong, Chaojun Xiao, Cunchao Tu, T. Zhang, Zhiyuan Liu, and Maosong Sun. 2019. Jec-qa: A legal-domain question answering dataset. *AAAI Conference on Artificial Intelligence*.

A Appendix : Wechat Screenshots Evidence Legal Processing Task

Workflow of Legal Practitioners

Legal professionals—especially plaintiff-side attorneys—follow a structured yet iterative workflow when preparing WeChat screenshots for courtroom use. As illustrated in Figure 3, this process typically unfolds across five interrelated stages: (1) preliminary screening, (2) timeline reconstruction, (3) key information extraction, (4) legal grounding, and (5) evidence cataloging and strategy formulation. These steps often involve back-and-forth revision

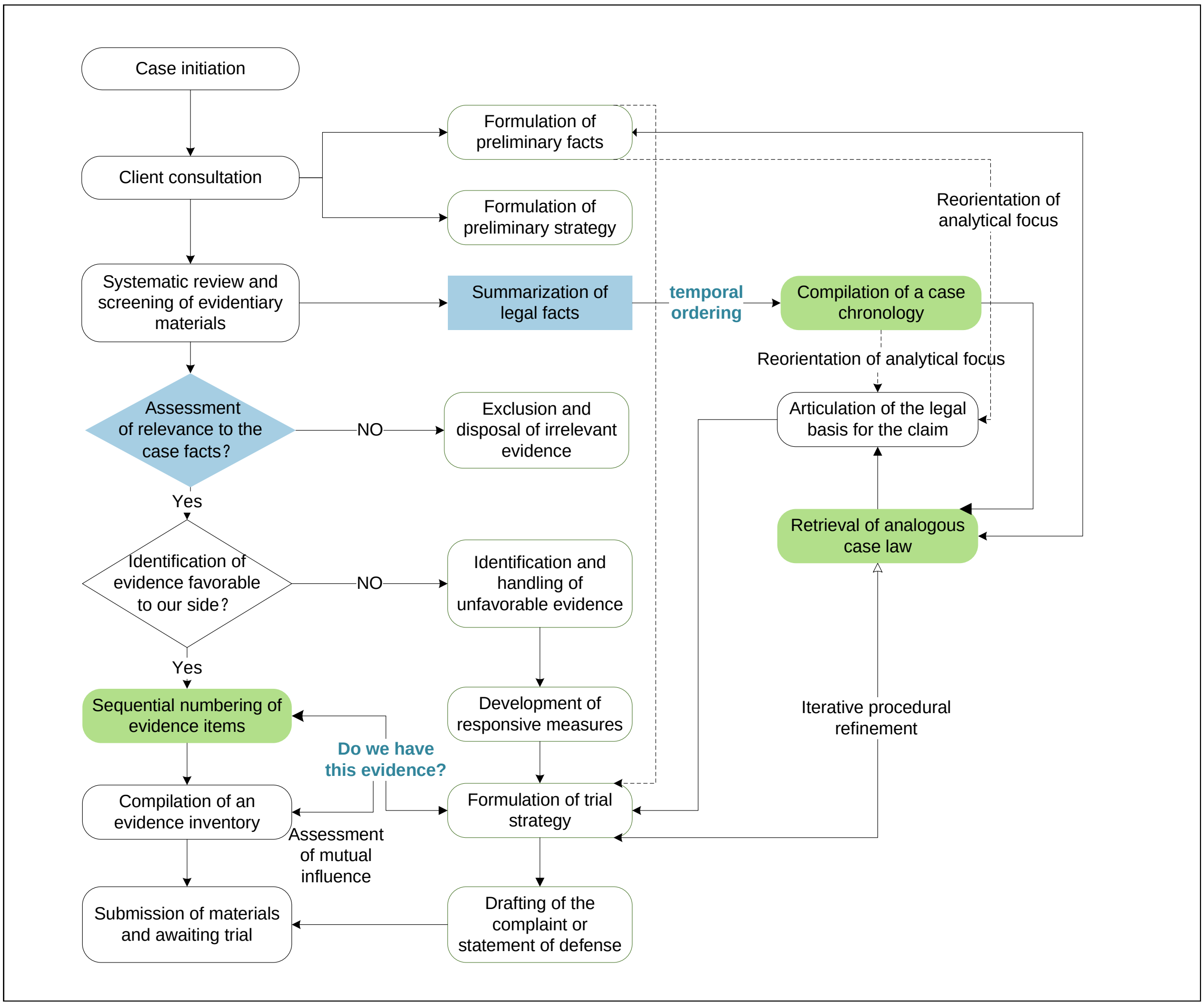


Figure 3: Workflow of legal practitioners when processing WeChat screenshots as evidence. **Blue components** indicate stages covered by our dataset—including structured KIE and VQA. **Green components** represent downstream legal reasoning goals such as timeline reconstruction, dispute grounding, and strategy formulation, which build upon the outputs of our benchmark. This alignment illustrates how SCREENSHOTLEGALBENCH supports core stages of real-world evidentiary workflows.

as new information emerges or legal interpretations are refined.

ScreenshotLegalBench mirrors this legal workflow by decomposing the Legal Screenshot Evidence Understanding (LSEU) task into modular components. Specifically:

- **Stages 1–3** are supported by the *KIE* task and local *VQA*, which detect speaker turns, extract transaction details, and identify message timestamps.
- **Stage 4** is operationalized as the `case_text` generation task, where models synthesize multiple screenshots to infer the dispute's legal basis.

As shown in Figure 3, the step of "summarization of legal facts" corresponds directly to our case reasoning module. This process is annotated in our dataset, though for evaluation purposes we adopt a simplified scheme: rather than scoring long-form legal texts, we assess whether the predicted output includes correct cause-of-action categories.

Visual-Legal Mapping of Evidence Elements

To enable automated legal understanding, ScreenshotLegalBench captures a range of heterogeneous visual elements in chat screenshots and aligns them with their legal semantics:

Avatars and nicknames provide identity signals for speaker verification. **Timestamps** serve as anchors for timeline reconstruction and causal ordering. **Chat bubbles and textual content** carry the core of intent and factual statements. **Image blocks** may represent either expressive content or direct legal exhibits. **Transfer notifications** frequently denote contractual performance or financial disputes. **Files and attachments** often indicate delivery obligations in cooperative agreements. **Emojis and quoted speech** encode attitudes, denials, or acknowledgments, which can be crucial for interpreting legal intent.

By aligning such fragmented visual-textual content with structured legal interpretation, ScreenshotLegalBench establishes a tractable framework for multimodal legal AI—grounded in the actual evidentiary workflows of judicial practice.

B Appendix : Dataset Implementation Details

B.1 Annotation Schema

As shown in Table 5, ScreenshotLegalBench adopts a five-level hierarchical annotation schema designed to meet the diverse demands of multimodal legal tasks. This schema integrates visual layout, semantic content, and legal reasoning to support both structured extraction and high-level judicial analysis.

Level 1 annotates global attributes of each screenshot, including legality, chat type, and case category, to facilitate filtering and legal classification. Level 2 focuses on layout elements such as message bubbles, avatars, and timestamps, using bounding boxes to support object detection and automated parsing. To reduce annotation cost while preserving effectiveness, full layout annotations are only provided for 50 screenshots. Level 3 structures message-level fields including timestamp, content, and speaker, supporting downstream dialogue reconstruction and evidence linkage. Level 4 further normalizes and formats semantic fields, such as monetary transfers and file metadata, ensuring compatibility with legal expression standards. Level 5 introduces VQA annotations, targeting both global and local reasoning about legal validity, intent, and evidentiary value (see Table 6 for examples).

This layered design ensures ScreenshotLegalBench supports both low-level structure-aware pretraining and high-level legal understanding, making it suitable for retrieval, reasoning, and structured generation tasks.

B.2 Object Detection For Sceenshots Layout Training

We adopt DETR as the screenshot layout detector, fine-tuned on our augmented WeChat chat screenshot dataset to detect message bubbles, avatars, timestamps, and other UI elements.Training configuration detail in Table 7

B.3 Timestamp Imputation

To enrich the temporal context of screenshots lacking explicit timestamps, we introduce a sliding-window-based imputation mechanism at the screenshot level. Considering real-world scenarios where multiple conversations may coexist and screenshot order can be disrupted, we first perform session-level clustering using OCR-extracted chat titles, followed by intra-session timestamp sorting and imputation.

Screenshots are categorized into three types based on the presence of timestamps: (1) If a single timestamp is detected, it is directly assigned as

Field Type	Attribute Name	Annotation Detail	Level	Annotation Format
Global Properties	Screenshot Validity	Whether it is a standardized chat screenshot	Level 1	Enumeration
	Chat Type	Group or private conversation		Enumeration
	Legal Relevance	Whether the screenshot has legal implications		Enumeration
	Case Type	Preliminary case classification (e.g., loan, contract)		Enumeration
Layout Elements	Avatar	`avatar_bbox`	Level 2	Bounding Box
	Message Bubble	`message_bbox`		Bounding Box + Text
	Chat Title / Group Name	`header_bbox`		Bounding Box
	Nickname	`nickname_bbox`		Bounding Box + Text
	Timestamp Region	`timestamp_bbox`		Bounding Box + Text
	Transfer Block	`transfer_bbox`		Bounding Box + Text
	File Block	`file_bbox`		Bounding Box + Text
	Image Block	`image_bbox`		Bounding Box + Category
	Emoji / Meme	`meme_bbox`		Bounding Box + Text
	Voice Message	`voice_bbox`		Bounding Box
	Recall Prompt	`withdraw_bbox`		Bounding Box
	Translation Block	`translate_bbox`		Bounding Box + Text
	Quote / Comment	`comment_bbox`		Bounding Box + Text
	Failed Message	`unpassed_message`		Bounding Box + Text
	Other Elements	`other`		Bounding Box
Message Fields	Speaker	`speaker`	Level 3	Enumeration / String
	Message Time	`timestamp`		Time String
	Message Content	`content`		Raw Text
Semantic Fields	Transfer Info	`transfer`	Level 4	Normalized String
	File Name and Type	`file`		Normalized String
	Image Description	`image`		Generated Text
	Emoji Polarity	`meme`		Enumeration
Legal QA Fields	Intent Analysis	"What intent is expressed in this message?"	Level 5	Text QA
	Legal Reasoning	"Please analyze the legal implications of this situation."		Text QA
	Transfer Nature	"What is the legal nature of the received transfer?"		Text QA
	File Legality	"Is the sent file direct legal evidence?"		Text QA

Table 5: Hierarchical annotation schema of ScreenshotLegalBench, covering five levels from global classification to legal reasoning.

the screenshot's temporal feature. (2) For multiple timestamps, we apply a heuristic to assign each timestamp to subsequent messages and compute the screenshot's time value as the arithmetic mean of all detected timestamps. (3) For screenshots with no timestamp, we estimate the time feature based on its position in the session sequence via a sliding average of neighboring screenshots.

Formally, let $t_{i,j}$ denote the imputed time for the j-th screenshot in the i-th session. Its value is computed as:

$$t_{i,j} = \frac{1}{k} \sum_{l=1}^{k} t_{i,j-l} \tag{1}$$

where k is the sliding window size, controlling how many preceding screenshots contribute to the estimation. This process is performed within each conversation thread, and the resulting timestamp is propagated to all messages in the corresponding screenshot for downstream context modeling and temporal reasoning.

For tasks requiring global temporal order (e.g., event timeline reconstruction), all screenshots can be sorted directly without regard to session boundaries. For tasks that depend on conversational structure, timestamp estimation and ordering are maintained per session.

Importantly, this imputation strategy is based on an engineering assumption that screenshot order roughly reflects message chronology. While this generally holds in user-submitted datasets, it may be invalid in legal contexts involving manipulation or reordering. Therefore, this method is positioned

Element Type	Example Question	Annotation Goal
Message Text	What is the intent of this message?	Identify expressions of intent (e.g., promise, request, warning)
	Who is the speaker?	Match speaker name for identity tracking
	What legal issue may be implied?	Perform legal inference (e.g., breach, infringement)
Avatar and Nickname	Is the avatar consistent with the nickname?	Verify identity coherence
Transaction Record	Who are the sender and recipient?	Determine transaction direction
	What is the legal nature of this transfer?	Classify as donation, payment, etc.
	What is the amount transferred?	Record monetary value (0 if unreadable)
Quoted Content	What is the speaker's attitude toward the quote?	Distinguish affirmation, denial, or doubt
Emoji	Does the emoji express affirmation or negation?	Interpret sentiment or intention
	What emotion is conveyed?	Provide cultural interpretation (e.g., sarcasm)
Dialogue Name	Does the name reflect identity?	Link to legally relevant identity info
	What is the legal relationship between parties?	Infer from context (e.g., employer–employee)
Timestamp	What is the message time?	Support timeline reconstruction
Other Elements	Was this message recalled?	Judge evidentiary validity
	Is the speech-to-text reliable?	Assess transcript usability
	What file was sent?	Record name, type, and purpose

Table 6: Examples of local legal reasoning questions in the VQA task. This set is under annotation and not yet released or evaluated.

Setting	Value
Model	DETR (COCO-pretrained)
Data Augmentation	Brightness, Crop, Flip, Rotate
Anchor Generation	k-means clustering + elbow method
Optimizer	AdamW
Learning Rate	5×10^{-5}
Batch Size	8
Epochs	50
LR Scheduler	Cosine Annealing
Early Stopping	Based on validation mAP
Metrics	mAP, IoU

Table 7: Training configuration for DETR-based screenshot layout detection.

as a heuristic for enhancing contextual coherence, not for evidentiary authentication or precise legal timeline reconstruction. Future work may incorporate device metadata or cross-image logical cues to improve legal robustness and applicability.

Note on Scope. While the timestamp imputation strategy described here is designed to support multi-image temporal modeling—especially for future tasks involving conversation reconstruction or inter-message reasoning—our current benchmark evaluation remains screenshot-level, with each VQA or KIE instance based on a single image input. This section primarily serves to document the semi-automatic annotation and reasoning methods applied during partial KIE labeling. It lays the groundwork for subsequent extensions of ScreenshotLegalBench toward multi-screenshot and temporally-aware legal understanding benchmarks.

B.4 JSON Schema

To support structured KIE from WeChat chat screenshots, we define a unified JSON output format that organizes each conversation into timestamped message entries with bounding box and semantic attributes. Figure 4 presents an example of the structured annotation used in KIE tasks.

B.5 Annotation Interface UI for VQA Tasks

To facilitate structured annotation for the VQA tasks in ScreenshotLegalBench, we employed the LabelU platform to design a dual-level labeling interface. The annotation process includes both global-level and local-level legal reasoning questions.

Global questions focus on the legal attributes of the entire screenshot—e.g., whether it constitutes valid evidence or what type of legal dispute it may relate to. Local questions target specific elements within the screenshot, such as a particular message, emoji, or transaction, and aim to elicit fine-grained legal interpretations.

Figure 5 shows an example of a global question annotation scenario, where the screenshot is assessed for its potential relation to a partnership dispute. Figure 6 displays a local question focused on a transfer message, prompting the annotator to

Figure 4: Example Data Instance for Annotation

determine its legal nature.

C Appendix : Datasets statistics

Object Detection Subset. This subset contains 50 manually annotated screenshots with a total of 945 bounding boxes across 15 interface element categories, used to train the layout detection models. As shown in Table 7, the majority of bounding boxes are concentrated in `message` and `avatar` regions, reflecting the visual dominance of conversation bubbles and speaker identity in chat interfaces.

KIE Training Set. The training set contains 39,477 message units extracted from screenshots via a semi-automatic pipeline. It includes 145,044 structured field annotations. As shown in Table 8, all samples have both `speaker` and `message_bbox`, while 85.5% include `content`, and 59.3% contain timestamp information. Additionally, the dataset captures non-textual legal indicators such as `transfer` and `image`.

KIE Evaluation Set. This subset consists of 143 human-annotated screenshots comprising 696 message units and 2,678 structured fields. Table 8 summarizes the distribution. Most messages include `speaker`, `message_bbox`, and `content`. Although rarer, legal fields such as `transfer`, `image`, and `file` are included due to their evidentiary value.

VQA Subset. The VQA set includes 1,176 chat screenshots annotated for multiple legal understanding tasks. As shown in Table 9, 38.9% are considered valid legal dialogs, and the same percentage were judged as evidential. A total of 502 samples include a textual case analysis written by legal professionals. Due to class imbalance, chat type (private vs. group) is only used as an auxiliary label.

D Appendix : Evaluation Detail

We evaluate a diverse set of baseline approaches on ScreenshotLegalBench to establish performance benchmarks for both KIE and VQA tasks. In this section, we describe the experimental setup, baseline methods, and implementation details.

D.1 Evaluation Results

Table 10 presents the performance of baseline and fine-tuned models on the KIE task, evaluated over the private chat subset of ScreenshotLegalBench. Scores reflect structured output quality in terms of format validity, spatial alignment (IoU), and content accuracy. Fine-tuned InternVL2.5-2B achieves the highest overall score of 0.6921, demonstrating strong improvements across all dimensions.

Table 11 reports the performance of baseline models on two classification sub-tasks: (1) *Classify*, which determines whether a screenshot qualifies as legally meaningful chat evidence, and (2) *Evidence*, which assesses whether the screenshot conforms to a valid legal format. All results are based on a unified evaluation setting using non-structured prompts. Across both tasks, most models exhibit limited performance, with low F1 scores and high variance across metrics. Notably, Qwen2.5-VL-3B-Instruct(Bai et al., 2025) achieves relatively higher classification accuracy, while Qwen2.5-VL-72B(Bai et al., 2025) shows better recall for evidence detection, albeit with poor precision.

Table 12 summarizes the results for the third subtask: *case_text*, which requires generating a plausible legal cause of action from multi-image inputs. We evaluate models using a weighted composite score that aggregates hit rate (i.e., dispute match), semantic similarity, and length alignment. Among all tested models, Qwen2.5-VL-32B-Instruct outperforms others, followed by Qwen2.5-VL-3B-Instruct, indicating the benefit of larger model scales and instruction tuning. Nevertheless, overall scores remain modest, suggesting that multi-image legal reasoning remains a challenging task for current VLMs.

D.2 Field Validation Rules for KIE Tasks

Each predicted message is considered structurally valid only if it contains the fields `speaker`, `timestamp`, `content`, and `message_bbox`. The `timestamp` must include at least one digit and pass regex-based sanity checks. Bounding boxes must be well-formed 4-tuples with positive width and height. For model outputs in invalid JSON or partial structures, we apply a fallback parser with bracket completion and nested field recovery. Messages failing all checks are excluded from scoring.

D.3 Finetune Configuration

We fine-tune the InternVL2(Chen et al., 2024b) and InternVL2.5(Chen et al., 2024a) models using the QLoRA approach; key training hyperparameters are listed in Table 13.

Figure 5: Annotation interface for global-level legal VQA tasks. This example shows a screenshot being annotated for its potential connection to a partnership dispute.

Field	Train	Eval	Struct.	Completeness	BBox	Time Validity	Content	Metric
speaker	39477	696	✓	✓	✗	✗	✓	TP / Total
message_bbox	39477	696	✓	✓	✓	✗	✓	IoU
content	33753	681	✓	✓	✗	✗	✓	λ SeqSim + (1-λ) LCS
timestamp	23397	522	✓	✓	✗	✓	✓	F1 (digit-check)
dialog_name	3208	61	✓	✓	✗	✗	✓	λ SeqSim + (1-λ) LCS
image	2661	10	✓	✓	✗	✗	✓	λ SeqSim + (1-λ) LCS
transfer	3071	10	✓	✓	✗	✗	✓	λ SeqSim + (1-λ) LCS
file	–	2	✓	✓	✗	✗	✓	λ SeqSim + (1-λ) LCS

Table 8: Summary of annotated fields across KIE dataset subsets and evaluation criteria.

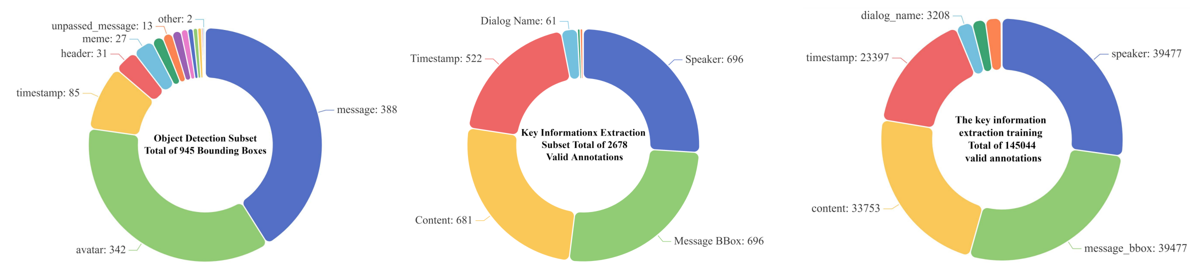
D.4 Prompt Templates

The following prompt templates were used during evaluation. Figure 8 shows the full Chinese prompt used for zero-shot evaluation. The first line of the template (“Please extract structured information from this chat screenshot”) was also used as the fine-tuning instruction.

D.5 Ablation Results

To understand the impact of structural signals, we ablate the use of KIE-enhanced prompts in VQA (Table 14) and the effect of bounding-box inputs in KIE (Table 15).

Figure 6: Annotation interface for a local-level legal VQA task. The annotator is asked: “What is the legal nature of this transfer?”



(a) Proportion of bounding boxes for each category in the object-detection subset (total = 945)

(b) Distribution of non-empty field annotations in the KIE evaluation subset (total = 2,678 entries)

(c) Distribution of non-empty field annotations in the KIE training set (total = 145,044 entries)

Figure 7: Annotation distribution statistics across the object detection and KIE subsets of ScreenshotLegalBench. Left: bounding box category ratio (N=945). Center: non-empty field counts in the KIE evaluation set (N=2,678). Right: non-empty field counts in the KIE training set (N=145,044).

Task Dimension	Label	Count	Percentage	Total
Dialog Type	legal_dialog	457	38.9%	1,176
	nonlegal_dialog	664	56.5%	
	not_dialog_but_legal	21	1.8%	
	not_dialog_and_nonlegal	34	2.9%	
Evidence Validity	is_evidence	457	38.9%	1,176
	not_evidence	719	61.1%	
Chat Type	private_chat	1,142	97.1%	1,176
	group_chat	34	2.9%	
Case Reasoning	with_case_text	502	42.7%	1,176
	without_case_text	674	57.3%	

Table 9: Annotation distribution in the VQA subset (total = 1,176).

Model	Format Score	IoU	Content Score	Overall Score	# Valid Samples
InternVL2-2B	0.9131	0.0009	0.3260	0.6195	143
InternVL2-2B (fine-tuned)	0.9517	0.0603	0.3909	0.6713	143
InternVL2.5-2B	0.9764	0.0006	0.2839	0.6302	143
InternVL2.5-2B-ft (run-13)	0.9644	0.0420	0.3944	0.6794	143
InternVL2.5-2B-ft (run-14)	0.9472	**0.1044**	**0.4369**	**0.6921**	143
Qwen2-VL-2B-instruct	0.3496	0.0006	0.0617	0.2057	143
Qwen2-VL-7B-instruct	0.5806	0.0000	0.2064	0.3935	31
Qwen2.5-VL-3B-instruct	0.9355	0.0012	0.2293	0.5824	31
Qwen2.5-VL-7B-instruct	0.0398	0.0009	0.0151	0.0274	143

Table 10: Evaluation results on the ScreenshotLegalBench KIE task (private chat subset). All scores are averaged over valid samples with parsing.

Model	Classify Task				Evidence Task			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Qwen2.5-VL-32B	0.0000	0.0000	0.0000	0.0000	0.0382	0.0144	0.3333	0.0275
Qwen2.5-VL-72B	0.0357	0.0556	0.0238	0.0333	0.0328	0.0242	0.4444	0.0447
Qwen2-VL-7B-Instruct	0.0513	0.0635	0.0286	0.0394	0.0000	0.0000	0.0000	0.0000
Qwen2.5-VL-7B-Instruct	0.0357	0.0536	0.0204	0.0296	0.0244	0.0240	0.3606	0.0404
InternVL2-2B	0.0123	0.0167	0.0048	0.0074	0.0182	0.0007	0.0095	0.0014
InternVL2.5-2B	0.0602	0.0510	0.0340	0.0408	0.0113	0.0017	0.0069	0.0019
Qwen2-VL-2B-Instruct	0.0120	0.0159	0.0071	0.0099	0.0026	0.0003	0.0010	0.0004
Qwen2.5-VL-3B-Instruct	0.2143	0.0310	0.0857	0.0456	0.0023	0.0003	0.0016	0.0005

Table 11: Classification and Evidence Evaluation Results (w/o Structured Prompt)

Model	Hit Rate	Similarity	Length Score	Weighted Score
InternVL2-2B	0.0253	0.0233	0.1169	0.0430
InternVL2.5-2B	0.0253	0.0415	0.1278	0.0507
Qwen2-VL-2B-Instruct	0.0633	0.0200	0.0647	0.0506
Qwen2-VL-7B-Instruct	0.0443	0.0376	0.1088	0.0552
Qwen2.5-VL-3B-Instruct	0.1329	0.0333	0.1127	0.0990
Qwen2.5-VL-32B-Instruct	**0.2089**	0.0187	0.0537	**0.1208**
Qwen2.5-VL-72B-Instruct	0.1250	0.0301	0.0566	0.0829

Table 12: Cause-of-action generation performance (multi-image reasoning).

<image>\n请从这张聊天截图中提取结构化信息,
Format: ```
{
\"dialog_name\": \"<对话名称>\",
\"conversation\": [{
\"timestamp\": \"<第一条消息的时间戳>\",
\"speaker\": \
"<如果是右侧发出则是avator_0, 左侧发出是avator_1>\",
\"content\": \"<说话内容>\",
\"message_bbox\": {
\"min_x\": <边界框的min_x, 是个数字, 如917>,
\"max_x\": <边界框的max_x, 是个数字>,
\"min_y\": <边界框的min_y, 是个数字>,
\"max_y\": <边界框的max_y, 是个数字>
},
\"image\": \
"<如果是图片或表情包, 需要描述这个图片>\",
\"transfer\": \
"<如果有转账信息, 填写在这里, 如微信转账请收款¥520.00>\",
\"file\": \
"<如果有excel或者doc等文件, 把文件名填写在这里, 记得带上文件格式后缀, 如2004年度数据分析表.xlsx>\"
},
...
]
}
```

**&lt;image&gt;\nPlease extract structured information from this chat screenshot.**
Format: ```
{
"dialog_name": "&lt;Name of the conversation&gt;",
"conversation": [{
"timestamp": "&lt;Timestamp of the first message&gt;",
"speaker": "&lt;If the message is on the right side, use avator_0; if on the left side, use avator_1&gt;",
"content": "&lt;Text content of the message&gt;",
"message_bbox": {
"min_x": &lt;Minimum x-coordinate of the message bounding box (e.g., 917)&gt;,
"max_x": &lt;Maximum x-coordinate of the message bounding box&gt;,
"min_y": &lt;Minimum y-coordinate of the message bounding box&gt;,
"max_y": &lt;Maximum y-coordinate of the message bounding box&gt;
},
"image": "&lt;If the message contains an image or emoji, provide a description of the visual content&gt;",
"transfer": "&lt;If the message contains a transaction, describe it here, e.g., WeChat Transfer: Please confirm receipt ¥520.00&gt;",
"file": "&lt;If the message includes a document (e.g., Excel or Word), write the filename including its extension, e.g., 2004年度数据分析表.xlsx&gt;"
},
...
]
}
```

Figure 8: Chinese prompt used during KIE pass@1 evaluation (left), with English translation shown on the right.

Hyperparameter	**Value**
Max sequence length	8192
Batch size (per GPU)	1
Gradient accumulation	2
Epochs	1 / 4
Optimizer	AdamW
LR scheduler	Cosine decay
Warmup ratio	3%
Initial learning rate	5×10^{-5} / 1×10^{-4}
LoRA rank	16
LoRA scaling factor	16
LoRA dropout	0.05
Gradient clipping	1.0
Layer-wise LR decay	0.75

Table 13: Fine-tuning hyperparameter configuration.

Table 14: Ablation: Effect of structured KIE input on VQA tasks.

Setting	Task	Accuracy	Macro P	Macro R	F1 Score
w/o KIE guidance	classify	0.4286	0.3333	0.1428	0.1999
+KIE-augmented input	classify	**0.8333**	**0.5000**	**0.4167**	**0.4545**
w/o KIE guidance	evidence	0.0157	0.0013	0.0107	0.0019
+KIE-augmented input	evidence	**0.0187**	**0.0126**	**0.0606**	**0.0204**

Table 15: Performance of InternVL2-2B on ScreenshotLegalBench with and without `message_box` bounding boxes

Model	**Format Score**	**Avg. IoU**	**Content Score**	**Overall Score**
InternVL2-2B (w/ bbox)	0.9384	0.0338	0.3703	0.6544
InternVL2-2B (w/o bbox)	0.7579	0	0.3934	0.5756