

1 Appendix

2 We provide more details of the proposed method and additional experimental results to help better understand our paper. The appendix is
3 organized to present comprehensive information on our experimental setup, prompting strategies, additional results, and limitations of the
4 current approach.

5 Contents

6	A	Experiment Setting	1
7	A.1	Datasets for Evaluation.	1
8	A.2	Baselines	1
9	B	Additional Experimental Results	2
10	B.1	Comparative Analysis of OCR Methods	2
11	B.2	CE Thresholds Analysis	3
12	B.3	Model Calibration Analysis	4
13	B.4	Identical Model Sampling Analysis	4
14	B.5	GPT-4o Output Verbosity Analysis	5
15	B.6	Distance Metric Comparison	5
16	C	RealData Source	6
17	D	Annotation Interface for Human Evaluation	7
18	E	Algorithm Details	9
19	E.1	CE Hyperparameters and Settings	9
20	F	Computational Cost and Efficiency Analysis	10
21	G	OCR Evaluation Case Studies: CE vs. VLM-as-Judge Comparison	11
22	H	Prompt	13
23	H.1	Evaluation Metrics	14
24	H.2	Details of Compared Methods	14
25	I	Ensemble Results	14
26	I.1	CCOCR	14
27	I.2	OCRBench-V2	14
28	I.3	OCRBench	15
29	J	Additional Representation Space Analysis	27
30	K	Limitations and Future Directions	28
31		References	28

32 A Experiment Setting

33 **Implementation.** We use inference-only evaluation with diverse VLMs following *vlmevalkit* [2] defaults. Most models run on single NVIDIA
34 H800 (80GB) GPU; largest models (e.g., Qwen2.5-VL-72B) use 2-4 GPUs. Our work involves **only inference, eliminating training and**
35 **minimizing computational cost.** Edit distance uses Levenshtein; cosine uses bge-m3 [1], applied only to OCRBench-V2. All experiments
36 apply temperature 0.0 (single-run) except Self-Consistency (T=0.7, 3 runs).

37 The experimental environment is configured with Ubuntu 22.04.2 LTS running on Linux kernel 5.10.134-16.103.al8.x86_64 with x86_64
38 architecture. PyTorch 2.6.0 is compiled with CUDA 12.4 support (+cu124), indicating compatibility with the installed CUDA toolkit.

39 A.1 Datasets for Evaluation.

40 Our research utilizes three datasets: OCRBench [3], OCRBench-V2 [3], and CCOCR [5], with the majority of our experiments conducted on
41 OCRBench.

42 To align our evaluation with human preferences in real-world contexts, we also curated a unique dataset consisting of 1,000 PDF pages
43 randomly selected from actual scenarios. These documents were processed using the state-of-the-art Qwen2.5-VL-72B model to perform
44 OCR tasks. The resultant texts were manually compared to the original PDF pages, and each was assigned a quality score ranging from
45 1.0 (Perfect Match) to 0.0 (Mostly Incorrect), with varying degrees of text matching accuracy. This diverse dataset collection allows for a
46 comprehensive assessment of OCR performance across various domains, languages, and formatting complexities, providing insights into
47 both machine efficiency and human-centric accuracy. Detailed scoring criteria and methodology are available in Appendix C. Data cases are
48 shown in Appendix G. Our annotated dataset is publicly available at: <https://huggingface.co/datasets/Aslan-mingye/OCR-Quality>.

49 A.2 Baselines

50 We establish two representative baseline approaches. These methods are widely adopted in current applications of VLMs, and serve as
51 comparative references across varying levels of complexity and supervision.

VLM-as-Judge. This baseline follows the mainstream paradigm of using large language models as evaluators to score candidate outputs. Specifically, we employ GPT4o¹, Qwen2-VL-72B, and Qwen2-VL-7B as evaluation models, each guided by standardized prompts to assess OCR output quality. The evaluation criteria emphasize semantic correctness, visual-text alignment, and structural fidelity.

Single Model Output. To better highlight the performance gains introduced by Consensus Entropy, we adopt a direct single-model output strategy as a baseline. In this setting, each image is processed by a single VLM (e.g., GPT4o or Qwen2.5-VL-72B), and its raw prediction is used as the final result without post-processing. This configuration reflects common practice in many real-world OCR applications and provides a conservative lower-bound reference for evaluating the effectiveness of our framework.

Self Consistency (SC). [4]selecting by majority voting among multiple samples from the same VLM. We set the temperature at 0.7 to make the models inference for 3 times and choose the best predictions.

B Additional Experimental Results

This section presents additional detailed analyses of our Consensus Entropy framework to complement the results discussed in the main paper. The extended results provide deeper insights into the behavior and effectiveness of different entropy calculation methods, the impact of distribution characteristics, and practical implications for deployment.

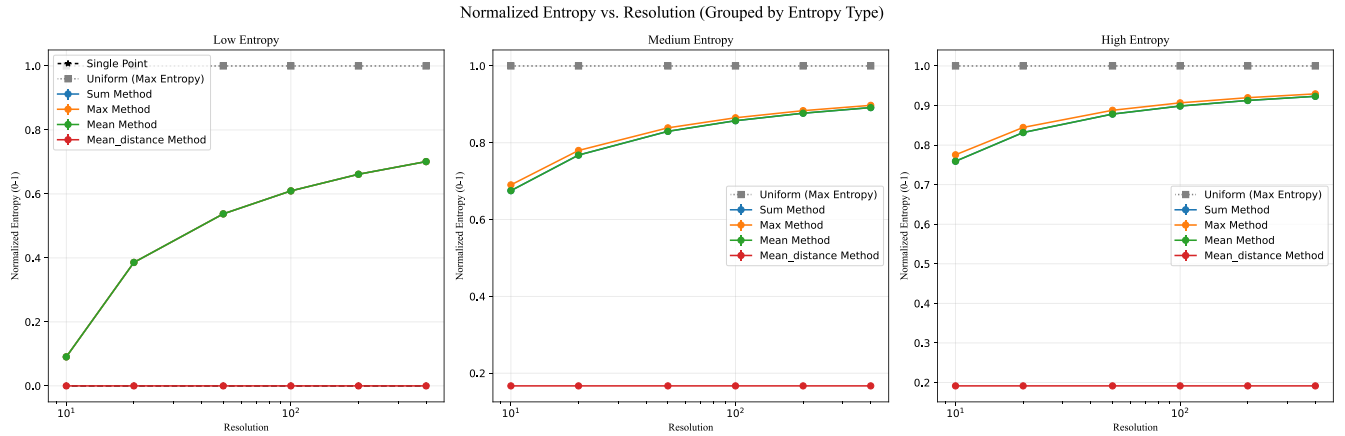


Figure 1: Normalized entropy comparison across different distribution types using the same aggregation methods. While Mean Distance shows the most distinctive pattern, all methods maintain similar relative positioning across distribution types, confirming the robustness of the entropy-based approach regardless of the specific computation method chosen.

Figure 1 extends our analysis of different aggregation methods by comparing their behavior across identical distribution types. The results demonstrate that while Mean Distance exhibits the most distinctive entropy pattern, all methods maintain consistent relative entropy values across distribution types. This consistency further validates our entropy-based approach by showing that the qualitative rankings of uncertainty remain stable regardless of the specific aggregation method employed.

B.1 Comparative Analysis of OCR Methods

To provide a comprehensive evaluation of different OCR approaches, we conducted a detailed comparison of three main strategies: Self-Consistency (SC@3), Routing-based ensemble, and Single model performance. Table 1 presents the results of this analysis, highlighting the relative performance improvements of each method.

The results demonstrate several key findings: First, the Routing-based ensemble approach consistently outperforms both SC@3 and Single model methods, with the best routing configuration achieving a score of 922, representing an 8.2% improvement over the best single model performance. Second, while SC@3 shows potential for improvement (2.7% over baseline when using Qwen2-VL-72B), its average performance actually decreases by 2.8%, indicating high variance in its effectiveness. Third, the gap between average and best performance is smallest for the Routing method (15 points) compared to SC@3 (47 points) and Single models (36 points), suggesting more consistent and reliable performance across different configurations.

Figure 2 investigates how kernel bandwidth settings and the correlation between distributions affect entropy estimates. A key finding is that higher correlation values consistently produce lower entropy curves, indicating that tightly clustered predictions (as would be expected for correct OCR outputs) naturally result in lower entropy measurements. This property is precisely what enables our CE metric to effectively distinguish between high-confidence and low-confidence predictions.

¹The specific version of GPT4o is gpt-4o-2024-11-20.

Table 1: Performance comparison of different OCR methods based on cumulative scores. SC@3 represents Self-Consistency with 3 samples, Routing refers to the ensemble with rephrasing, and Single indicates individual model performance. Blue highlights indicate the best performance in each category, while green highlights mark the second best.

Method	Score	Improvement (%)
SC@3 (Average)	828	-2.8
SC@3 (Best: Qwen2-VL-72B)	875	2.7
Routing (Average)	907	6.5
Routing (Best: Qwen2.5-VL-72B)	922	8.2
Single (Average)	852	-
Single (Best: Qwen2-VL-72B)	888	-

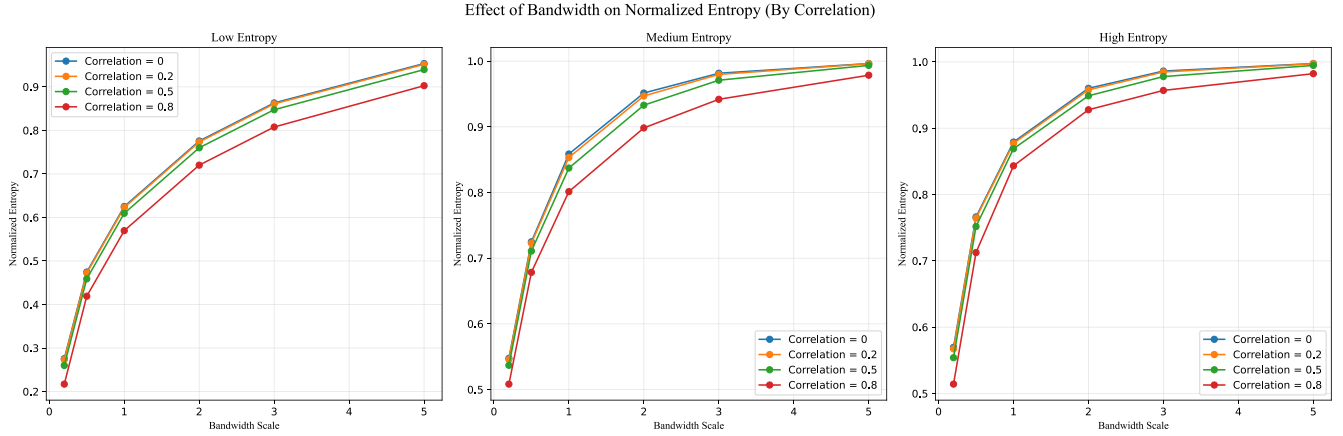


Figure 2: Effects of kernel bandwidth and distribution correlation on entropy estimates. Higher correlation values consistently result in lower entropy curves, highlighting how the spatial relationship between predictions influences the consensus measurement.

We also conducted experiments on OCRBench-V2 regarding the relationship between CE thresholds and benchmark scores, as shown in Table 7.

B.2 CE Thresholds Analysis

Table 7 reports the average scores of three models on OCRBench-V2 under different CE thresholds, showing that CE remains effective for verifying OCR and VQA-style tasks beyond the main OCRBench setting. Here, CE is used as a routing signal: lower thresholds keep only low-entropy (high-confidence) predictions, while higher thresholds allow more samples to pass.

To make the routing behaviour more explicit, we further summarize the accuracy-computation trade-off in terms of *accuracy improvement* vs. *rephrase ratio*. Table 2 (numbers reproduced from the main paper and rebuttal) shows, for representative models on OCRBench-V2, the gain in accuracy ($\Delta \times 10^2$) and the percentage of samples routed for rephrasing at different thresholds. Smaller θ values route more samples and yield higher accuracy but incur greater computation, while larger θ values are more efficient but less aggressive. For CE-OCR experiments on OCRBench, we use a single global threshold $\theta = 0.5$ (Section 4.3); for OCRBench-V2 routing, the sweep in Table 2 indicates that values around $\theta \approx 0.5$ offer a good balance between accuracy gains and routing cost.

Table 2: Threshold sensitivity analysis on OCRBench-V2. Each cell shows accuracy improvement ($\Delta \times 10^2$) / rephrase percentage (%). Lower θ routes more samples for rephrasing, improving accuracy at the cost of additional computation.

Model	$\theta = 0.95$	$\theta = 0.9$	$\theta = 0.8$	$\theta = 0.6$	$\theta = 0.2$
Gemini Pro	+1.6 / 11.5	+3.1 / 22.3	+4.3 / 43.5	+5.4 / 66.3	+6.1 / 91.5
GPT-4o	+3.8 / 15.0	+5.1 / 24.9	+6.8 / 48.7	+7.7 / 71.3	+8.8 / 91.2

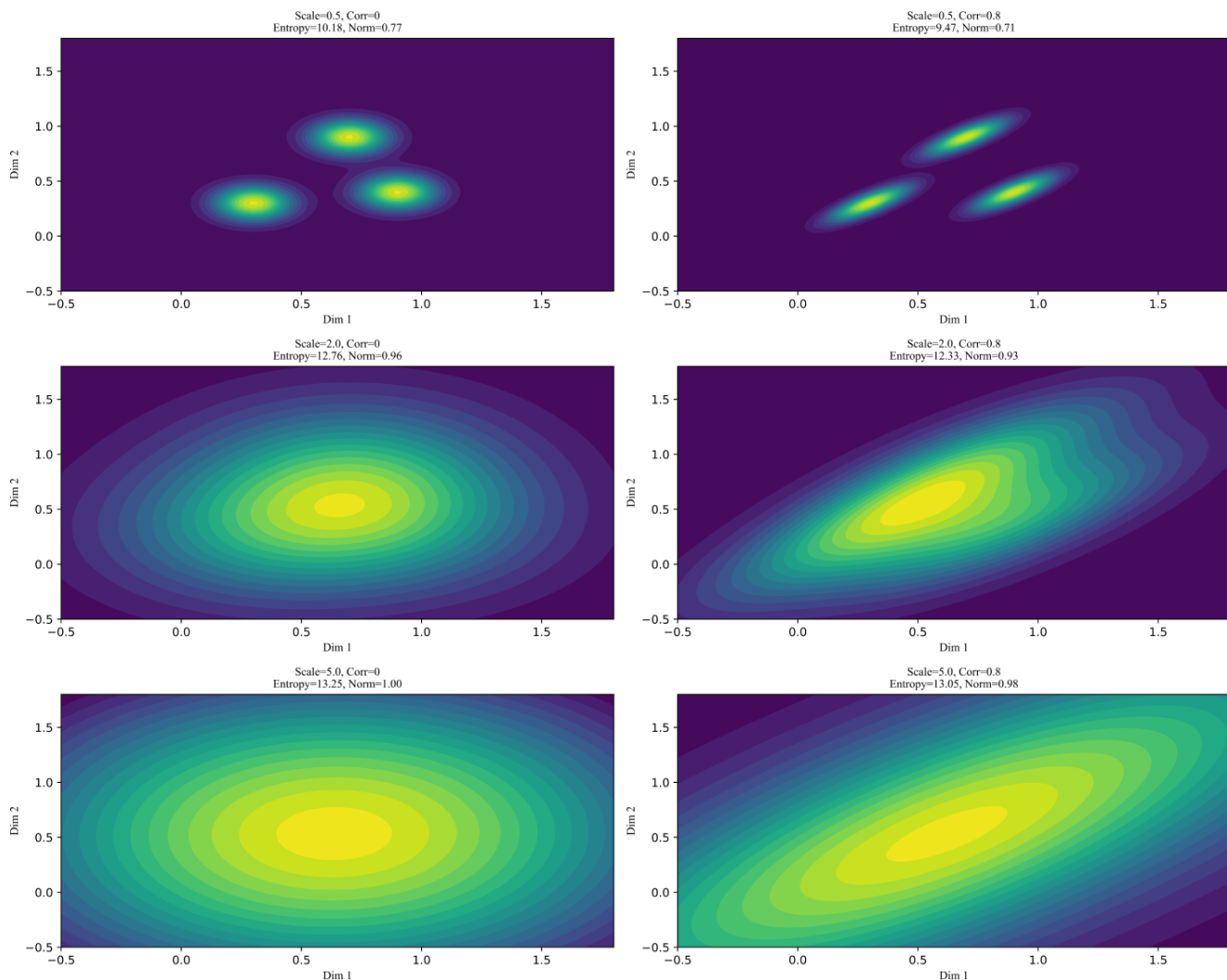


Figure 3: Visualization of entropy patterns across various scale and correlation settings. These examples illustrate how our entropy calculation responds to different distribution characteristics, with tighter, more correlated clusters resulting in lower entropy values regardless of scale.

95 B.3 Model Calibration Analysis

96 **Note:** ECE and Brier Score metrics are designed for classification probability calibration and are mismatched for our generation-based
 97 approach. **This analysis is provided for reference only and should not be used as the primary evaluation criterion for CE, which**
 98 **is a post-hoc uncertainty metric derived from output agreement rather than probability prediction.**

99 For completeness, we evaluated CE using Expected Calibration Error (ECE), Brier Score, and AUC metrics against VLM-as-Judge methods
 100 using Qwen2.5VL-72B with different reference model combinations (diverse: InternVL+GPT4o vs related: QwenVL series). Despite ECE and
 101 Brier scores being primarily designed for classification probabilities, CE achieves favorable metrics (best ECE: 0.0842, Brier: 0.1169, AUC:
 102 0.9226), though these values should not be overinterpreted given the methodological mismatch.

103 B.4 Identical Model Sampling Analysis

104 To evaluate CE’s robustness with minimal model diversity, we conducted experiments using identical models with stochastic sampling
 105 (temperature=0.7, 3 runs). Table 4 shows that even with the same model architecture and weights, CE-Ensemble consistently improves
 106 performance by leveraging sampling variance. This demonstrates that CE captures task-level uncertainty effectively, even when architectural
 107 diversity is absent.

Table 3: Calibration metrics for CE and VLM-as-Judge on Qwen2.5VL-72B predictions. CE demonstrates better calibration despite being designed for generation tasks rather than classification.

Method	ECE↓	Brier↓	AUC↑
Judge-Related	0.1090	0.1404	0.8041
Judge-Diverse	0.0952	0.1392	0.7727
CE-Related	0.0938	0.1362	0.9181
CE-Diverse	0.0842	0.1169	0.9226

Table 4: CE-Ensemble performance with identical models using stochastic sampling. Individual run scores shown in parentheses. CE consistently selects better outputs across sampling variance.

Model	Text Rec.	Scene VQA	Doc VQA	KIE	Formula	ALL
Qwen2VL-7B	272 (266,266,269)	173 (171,168,170)	155 (156,151,149)	185 (182,181,181)	67 (59,61,63)	852 (834,827,832)
Qwen2VL-72B	264 (265,263,261)	181 (179,182,177)	173 (166,170,173)	186 (183,183,183)	71 (61,64,71)	875 (854,865,865)
Qwen2.5VL-72B	264 (256,260,264)	176 (172,174,174)	175 (171,169,173)	184 (184,182,186)	71 (71,69,67)	870 (854,854,864)

B.5 GPT-4o Output Verbosity Analysis

During threshold analysis (Figure 4 in main paper), GPT-4o exhibited non-monotonic accuracy patterns. Investigation revealed that GPT-4o generates verbose outputs for short-answer questions in OCRBench. For example, when other models answer “11.90”, GPT-4o responds: “The total amount of this receipt is **RM 11.90**, as stated under ‘total incl. GST @6%’”. This verbosity artificially inflates CE values. Table 5 shows that after cleaning outputs via regex and LLM-based filtering, GPT-4o’s accuracy exhibits expected monotonicity with CE thresholds.

Table 5: GPT-4o performance under different CE thresholds before and after output cleaning. Cleaned outputs restore monotonic relationship between CE and accuracy.

CE Threshold	Original	Cleaned
≤0.7	0.781	0.831
≤0.4	0.833	0.884
≤0.2	0.833	0.928
≤0.1	0.844	0.930

B.6 Distance Metric Comparison

CE supports multiple distance metrics depending on task requirements. Table 6 compares Edit Distance (character-level) and Cosine Distance (semantic-level) across different task types. Edit Distance excels in pure OCR tasks requiring character-level precision, while Cosine Distance (using bge-m3) performs better on complex VQA tasks with semantic variations. Both maintain the same CE framework—only the pairwise similarity computation differs.

Table 6: Comparison of distance metrics for CE computation. The choice of metric depends on task characteristics: character-level precision vs semantic understanding.

Method	Granularity	Best For	Cost	Used In
Edit Distance	Character	OCR, Math, Code, Simple VQA	Very low (CPU)	Main OCRBench experiments; Table 1
Cosine Distance	Semantic	Complex VQA, Diverse questions	Low (small model)	Appendix OCRBench-V2; Table 4

Table 7: Performance of 3 models on OCRBench-V2 under Different CE Thresholds.

Model	CE Threshold	Total QAs	EN Overall	CN Overall	ALL Overall
Gemini-PRO	1.0	10000	0.519	0.431	0.520
	0.9	8816	0.549	0.465	0.546
	0.8	7252	0.580	0.535	0.582
	0.7	6009	0.614	0.560	0.610
	0.6	5221	0.643	0.588	0.633
	0.5	4499	0.640	0.596	0.652
	0.4	3276	0.702	0.650	0.668
	0.3	2323	0.772	0.698	0.701
	0.2	1664	0.785	0.648	0.743
	0.1	1041	0.828	0.690	0.790
GPT4o	1.0	10000	0.465	0.322	0.473
	0.9	8357	0.521	0.394	0.541
	0.8	6571	0.559	0.463	0.584
	0.7	5341	0.585	0.529	0.604
	0.6	4566	0.595	0.581	0.615
	0.5	3903	0.614	0.625	0.639
	0.4	3031	0.652	0.647	0.652
	0.3	2423	0.665	0.708	0.682
	0.2	1665	0.769	0.663	0.736
	0.1	1192	0.823	0.740	0.796
Intern2.5VL-26B	1.0	10000	0.494	0.442	0.532
	0.9	7160	0.558	0.511	0.580
	0.8	5948	0.581	0.553	0.593
	0.7	5198	0.614	0.551	0.620
	0.6	4629	0.626	0.623	0.642
	0.5	3976	0.634	0.646	0.664
	0.4	2850	0.673	0.630	0.697
	0.3	2056	0.677	0.658	0.744
	0.2	1638	0.776	0.696	0.772
	0.1	1118	0.829	0.742	0.799

118 C RealData Source

119 The datasets used in our OCR evaluation experiments were carefully curated to reflect diverse real-world document scenarios. Below we
120 present the detailed composition of these datasets, encompassing various languages, document types, and content domains.

Table 8: Sources of the human-labeled OCR Dataset

	ebook	paper	textbook	other
ZH	zhishilei, zhongwenzaixian, gift, thomas	-	by, kps, kmath, zhonggaokao, gaojiaoshe, gaodengjiaoyu, k12 edu platform, jiaoan, zju icles	-
EN	theeye, physicsandmathstutor, planetebook	escholarship, biorxiv, springer, sagepub, scholarworks, psyarxiv, chemrxiv, iop-science, royalsocietypublishing, criso	kps, bookboon, california 14sets, scholarworks	dev books repository
ML	renhang, banshujiang	-	openstax, math	-
Other	-	-	-	coursehero, studypool

121 The dataset composition encompasses a wide range of content types and sources, ensuring comprehensive evaluation of OCR capabilities.
122 Chinese language materials include electronic books from major repositories as well as educational content across various academic levels.

English content spans scholarly publications from multiple disciplines, educational materials, and literature. Multilingual resources incorporate specialized content with mixed language elements, while additional sources provide diverse formatting challenges. This heterogeneous collection enables robust assessment of OCR performance across domains, languages, and formatting complexities representative of real-world document processing requirements.

For human evaluation of OCR quality, we developed a standardized 4-level scoring system:

- (1) **Perfect Match (0.9-1.0)**: The prediction matches the image text exactly with no errors.
- (2) **Minor Errors (0.7-0.8)**: Very close to the image text with only small mistakes that do not affect understanding.
- (3) **Partially Correct (0.4-0.6)**: Contains noticeable errors or captures only part of the text.
- (4) **Mostly Incorrect (0.0-0.3)**: Largely incorrect or unrelated to the text in the image.

This scoring system was used consistently across all human evaluations in our experiments to ensure reliable assessment of OCR quality.

D Annotation Interface for Human Evaluation

To ensure the quality of our real-world dataset and provide reliable human judgments for OCR performance evaluation, we developed a dedicated annotation interface. Figure 4 presents our custom-built tool designed specifically for OCR evaluation tasks.

The annotation interface is implemented using Gradio, a Python library for creating web-based interfaces. This tool facilitates efficient human evaluation of OCR results by displaying the original image alongside the extracted text. Annotators can assign one of four quality ratings to each OCR output: (1) Completely Correct, (2) Minor Errors, (3) Partially Correct, or (4) Mostly Incorrect. This standardized rating system ensures consistency across evaluations and provides fine-grained assessment of OCR quality.

The interface includes several features to enhance annotation efficiency: a status display showing progress through the dataset, navigation controls to move between samples, the ability to skip difficult cases, and a direct jump function to specific image indices. All annotations are automatically saved to the original data file, creating a persistent record of human judgments. This annotation tool played a critical role in developing our ground truth evaluations and validating the effectiveness of the Consensus Entropy framework against human quality assessments.

Reload ./data.json and ./image

Image Preview

↓

OCR Recognition Content (LaTeX/Markdown source)

OMBRE--OMEN

northwest of Timor, from which the Ombay Pass, in the line of one of the best routes from Europe to China, separates it. It is about 900 square miles in extent, 65 miles long, and 15 broad, and presents a bold coast and lofty interior. The mountains are covered to their summits with lofty trees. It is inhabited by savage tribes, said to be fierce and treacherous, and carries on some trade with Timor in birds' nests and provisions, exchanged for iron-work, Chinese wares and linen, Allon or the northwest and Baliko on the southeast being the chief settlements and ports. It belongs to the Netherlands, and is included in the residency of Timor.

Ombre, òm'bër, a game of cards originating in Spain. It is usually played by three persons, with 40 cards (the eights, nines, and tens having been removed), and each player receives nine cards, three by three. The game is often mentioned in English 18th century literature.

Omdurman, òm-door'mán, Sudan, a native town on the White Nile, opposite Khartum. It was built as the capital of the Mahdi's successor, when Khartum was destroyed in 1885, and extends for four miles along the river bank. Under the Khalifa Abdallah it had a population of 500,000 inhabitants, living in one-storied mud huts, lining streets laid out on an orderly system, and guarded by a walled enclosure flanked with towers, accommodating 10,000 warriors. When the Khalifa was defeated, 2 Sept. 1898, Omdurman was hastily deserted, but with returning trade and prosperity has (1904) an estimated population of 60,000. Omdurman is the great mart of the gum-arabic, ivory, and ostrich-feather trade of northeastern Africa, and representatives of over thirty tribes congregate in its markets.

O'Meara, ò-má'ra, Barry Edward, English physician: b. Ireland 1778; d. London 3 June 1836. He was household physician to the Emperor Napoleon I. at Saint Helena, and published 'Napoleon in Exile' (1822). Originally a surgeon in the British navy, he was serving on the Bellerophon in that capacity 7 Aug. 1815 when Napoleon went on board. Napoleon noting O'Meara's skill and knowledge of Italian, desired the surgeon to accompany him to Saint Helena. Having obtained Admiral Keith's permission, O'Meara remained with the ex-emperor till July 1818. He was then recalled and deprived of his rank, for having accused Sir Hudson Lowe before the admiralty of cruel and arbitrary conduct.

OMEGA, ò-mé-gá or ò-míg'a, the last letter of the Greek alphabet: hence, figuratively speaking, the end or last of anything. See ALPHA.

Omen, a sign believed to prognosticate a future event, between which and the event foretold there appears no relation of cause and effect, but which is usually received as an intimation from a superior power. Omens have been common among most nations, and are often remembered and mentioned after they have ceased to be credited. Though generally classed among superstitions, they may sometimes be founded on some hidden relation in things, some natural law of sequence the ground of which is unknown. They have been chiefly in vogue in the ruder ages and communities, though under the name of auguries they retained their influence during the whole period of pagan antiquity, and through eminent warriors and other popular leaders in moments of extreme doubt and peril have given notable examples of faith in them. Sneezing was deemed ominous in the time of Homer, and Eustathius states that it was lucky or unlucky according as it was directed to the right or the left. Aristotle discusses the problem why sneezing from noon to midnight is good, and from midnight to noon bad. At noon it was propitious. Among the ancient Persians sneezing was esteemed fortunate, a sign of contest between the fiery soul and the earthly body, and of the victory of the former. When the emperor Monomotapa sneezes, says Codignus, it is proclaimed through the whole land as a signal for general joy. The itching of the nose implied that a stranger was coming. Burton, in his 'Anatomy of Melancholy,' states that 'to bleed three drops at the nose is an ill omen.' The spots on the finger nails were all ominous; the itching of the palm of the right hand promised a receipt of money; the doubling of the thumb within the hand was believed to have efficacy in avoiding approaching danger, and therefore the thumbs of dead persons were so folded. The way in which fires, candles, or lamps burned suggested divers omens. The superstition still prevails in many places that the howling of a dog by night presages a death in the neighborhood. Duncan Campbell expresses his faith in this omen, and adds: "Odd and unaccountable as it may seem, those animals scent death, even before it seizes a person." The screeching of the owl and the croaking of the raven have both in ancient and modern times been regarded as omens of some dire calamity. Divers presages concerning the weather have been derived from the habits of birds, bees, wasps, gnats, etc. Pennant states that many of the great families of Scotland received monitions of future events, especially of death, by spectres, wraiths, and shrieks. Fishermen and sailors discover omens in echoes, flashes, shadows, and other visible appearances. To throw a cat overboard, or lose a bucket, is believed to be unlucky. Whistling is supposed to stir up the wind. Stumbling has been the subject of numerous superstitions. Gaius Gracchus stumbled at his threshold on the morning of his death. To stumble on going out, says Bishop Hall, was mischievous; to stumble up stairs, says Grose, was lucky.

At the present day, in many parts of England and the United States, a superstitious belief in omens exists. It is regarded as unlucky to see first one magpie and then more; but two denote marriage or merriment; three, a successful journey; four, an unexpected piece of good news; five, that you will shortly be in a great company. To kill a magpie is to incur some terrible misfortune. When a person goes out on any important business, it is lucky to throw an old shoe after him. To present a knife, scissors, razor, or other sharp or cutting instrument to one's friend is unlucky, as they are apt to divide love and friendship. The falling of salt toward persons at table, the spilling of wine on their clothes, are evil omens. Breaking a looking glass betokens the death of the best friend of the person to whom it belonged. The burning of the cheeks, or tingling of the ears, that others were talking of us; if of the

Current Status

Index: 24 | 25 / 1000 | Not scored

1 Completely Correct

2 Minor Errors

3 Partially Correct

4 Mostly Incorrect

Previous

Skip

Next

Jump to index (not sequence number)

20

Jump

Figure 4: OCR Human Evaluation Interface: This annotation tool enables efficient assessment of OCR quality across multiple models. The interface displays the source image alongside OCR results, allowing annotators to assign quality scores based on a standardized 4-level rating system (from completely correct to mostly incorrect). The tool tracks progress, enables navigation between samples, and supports targeted image selection, facilitating comprehensive human evaluation of OCR performance on real-world data.

145 E Algorithm Details

146 This section provides detailed pseudocode for the key algorithms used in our Consensus Entropy framework.

Algorithm 1 Consensus Entropy for Multi-Model OCR Evaluation

Require: $\{O_1, O_2, \dots, O_n\}$ - outputs from n OCR models for image I
Require: σ - base kernel bandwidth parameter
Require: α - adaptive bandwidth scaling factor
Require: N - grid resolution for probability density estimation
 // Compute pairwise entropies and weights
for $i = 1$ to n **do**
 for $j = 1$ to $n, j \neq i$ **do**
 $\mathcal{E}_{ij} \leftarrow \text{ComputePairwiseEntropy}(O_i, O_j)$
end for
 $\bar{\mathcal{E}}_i \leftarrow \frac{1}{n-1} \sum_{j \neq i} \mathcal{E}_{ij}$ {Average entropy distance}
 $w_i \leftarrow \frac{1/\bar{\mathcal{E}}_i}{\sum_{k=1}^n 1/\bar{\mathcal{E}}_k}$ {Normalize weights}
 $\Sigma_i \leftarrow \sigma \cdot (1 + \alpha \cdot \bar{\mathcal{E}}_i) \cdot \mathbf{I}$ {Adaptive covariance}
end for
 // Project outputs to semantic space and estimate distribution on $N \times N$ grid
 $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\} \leftarrow \text{ProjectToSemanticSpace}(\{O_1, O_2, \dots, O_n\})$
 $P \leftarrow \text{EstimateDistributionOnGrid}(\{\mathbf{v}_i, w_i, \Sigma_i\}_{i=1}^n, N)$
 // Calculate entropy from $N \times N$ grid-based distribution
 $\delta \leftarrow -\sum_{x=1}^N \sum_{y=1}^N P[x, y] \cdot \log P[x, y]$ {Discrete entropy calculation on $N \times N$ grid}
return δ {Return Consensus Entropy}

Algorithm 2 CE-OCR: Entropy-guided Ensemble and Routing Framework

Require: I - input image containing text
Require: $\{M_1, M_2, \dots, M_n\}$ - set of OCR models
Require: M_{exp} - stronger vision-language model for rephrasing
Require: θ - entropy threshold for routing
 $\{O_1, O_2, \dots, O_n\} \leftarrow \{M_1(I), M_2(I), \dots, M_n(I)\}$ {Generate predictions}
 $\delta \leftarrow \text{ConsensusEntropy}(\{O_1, O_2, \dots, O_n\})$ {Calculate entropy using Algorithm 1}
if $\delta \leq \theta$ **then**
 // Use precomputed weights from Algorithm 1
 $O_{\text{final}} \leftarrow \text{WeightedEnsemble}(\{O_1, O_2, \dots, O_n\}, \{w_1, w_2, \dots, w_n\})$
else
 $O_{\text{ens}} \leftarrow \text{SimpleEnsemble}(\{O_1, O_2, \dots, O_n\})$ {Basic ensemble}
 $O_{\text{final}} \leftarrow M_{\text{exp}}(I, \{O_1, O_2, \dots, O_n\}, O_{\text{ens}})$ {Route to stronger model for rephrasing}
end if
return O_{final}

147 E.1 CE Hyperparameters and Settings

148 For reproducibility, we summarize the key hyperparameters used in all CE experiments.

149 **Distance Metrics.** As detailed in Table 6, CE supports two distance metrics: (1) character-level Edit Distance for pure OCR tasks,
 150 mathematical formulas, code, and simple VQA, and (2) semantic Cosine Distance for complex VQA and diverse questions. In practice, all
 151 OCRBench (main paper Table 1) and human-evaluation experiments use Edit Distance on raw text strings, while OCRBench-V2 experiments
 152 employ Cosine Distance based on bge-m3 embeddings (and bge-large-zh-v1.5 in early pilot experiments), following the configurations
 153 discussed in the rebuttal.

154 **Kernel and Grid Parameters.** In the semantic case, we apply isotropic Gaussian kernels with covariance $\Sigma_i = \sigma \cdot (1 + \alpha \cdot \bar{\mathcal{E}}_i) \cdot \mathbf{I}$ on a
 155 fixed $N \times N$ grid, as specified in Algorithm 1. The bandwidth base σ , scaling factor α , and grid resolution N are treated as global constants
 156 shared by all models on a given benchmark and tuned once on a held-out development split. We observe that CE is empirically robust to
 157 moderate changes in these values; performance is dominated by the choice of distance metric and routing threshold rather than fine-grained
 158 kernel tuning.

Sampling Temperature. Unless otherwise stated, all VLMs run with temperature 0.0 (greedy decoding). Self-Consistency baselines and identical-model CE-Ensemble experiments use temperature 0.7 with 3 samples, as reported in Section 4 and Table 4.

F Computational Cost and Efficiency Analysis

We measure both CE calculation costs and efficiency, comparing single-query verification (requiring one pairwise CE computation) against ensemble approaches.

Computational Cost of CE. We benchmarked CE computation on 1,000 OCR output pairs (average length ~1K characters). Table 9 presents detailed computational cost analysis. Edit Distance (CPU) is the fastest method at 0.0002s per computation, requiring no GPU resources. All methods except bge-m3 (CPU) achieve sub-0.1s computation times. For maximum efficiency, we recommend using Edit Distance for character-level tasks and GPU-accelerated models for semantic tasks, while avoiding CPU execution of large models.

VLM Cost Comparison. Paper "Figure 5: Performance comparison across token lengths" show that CE based on small VLMs **outperform stronger single models and require fewer tokens, less GPU memory, and offer faster inference**. Notably, while 70B+ models demand at least 4×80GB GPUs, multiple small models can run concurrently on a single GPU, highlighting the efficiency of the CE framework.

CE Method	Time Cost (s)			Memory (GiB)
	Avg	Med	Total	
Edit Distance (CPU)	0.0002	0.0000	0.1626	-
bge1.5-zh (GPU)	0.0116	0.0117	11.5716	1.19
bge-m3 (GPU)	0.0348	0.0311	34.7902	4.01
bge-m3 (CPU)	2.6518	2.0933	2651.7895	-

Table 9: Computational cost of CE on 1,000 long-text pairs (average length ~1K chars). Avg/Med/Total: average, median, total time cost of one pair.

Table 10: VLM cost comparison on OCRBench. Ensembles of small models achieve better performance with lower resource requirements than single large models.

Model	Time (s)	Token/s	Memory (GiB)
InternVL2.5-4B	409	6.98	7.90
InternVL2.5-8B	672	4.28	20.71
InternVL2.5-26B	951	3.03	52.67
InternVL2.5-78B	1772	1.62	158.53
Ovis2-4B	680	27.49	9.09
Ovis2-8B	611	30.27	18.26
Ovis2-34B	1191	15.33	66.60
Qwen2.5VL-3B	633	19.05	8.04
Qwen2.5VL-7B	771	153.16	16.99
Qwen2.5VL-72B	2715	7.52	140.98

171 **G OCR Evaluation Case Studies: CE vs. VLM-as-Judge Comparison**

172 To provide a more intuitive understanding of how Consensus Entropy (CE) compares with VLM-as-Judge methods in real-world OCR tasks,
173 we present several representative cases that highlight their respective performance characteristics.
174 **Scoring Methodology.** All cases use the 4-level scoring system defined in Appendix C: Perfect Match (0.9–1.0), Minor Errors (0.7–0.8),
175 Partially Correct (0.4–0.6), and Mostly Incorrect (0.0–0.3). For CE interpretation: low CE (< 0.3) indicates high consensus and likely correct
176 output; medium CE (0.3–0.7) suggests moderate disagreement; high CE (> 0.7) signals significant divergence and probable errors.

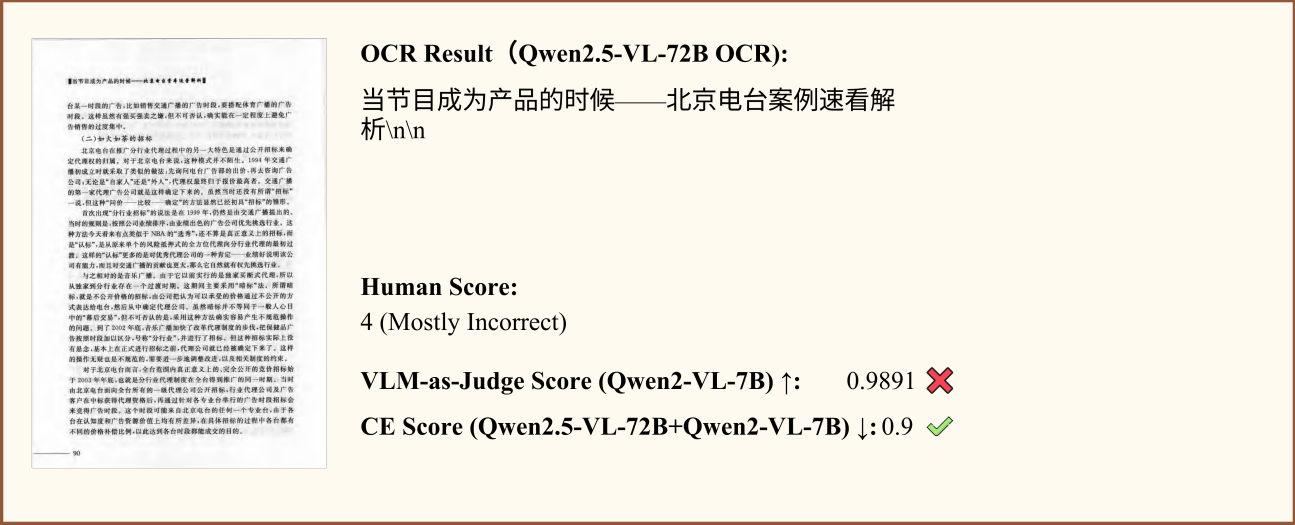


Figure 5: Case 1: High-quality Chinese OCR sample with strong agreement between human evaluation (4/4), CE score (0.9891), and VLM-as-Judge score (0.9).

177

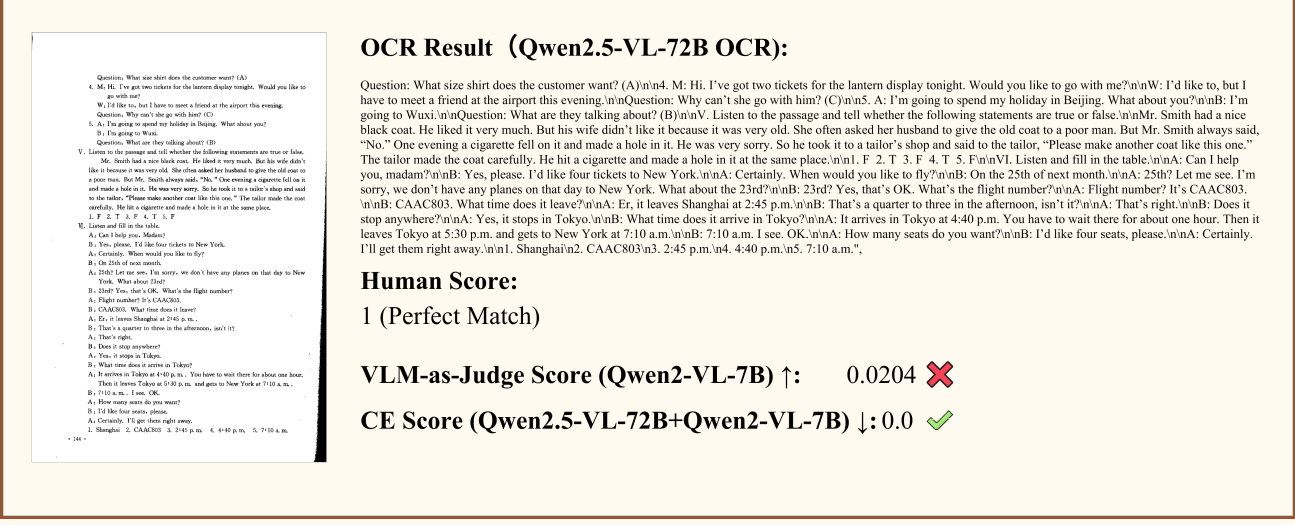
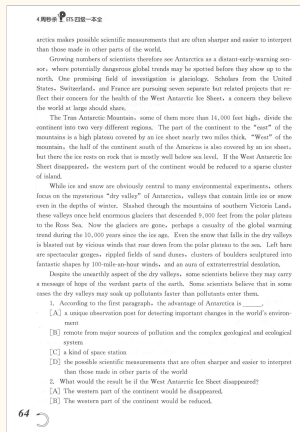


Figure 6: Case 2: Low-quality English OCR sample with unanimous poor quality assessment across human evaluation (1/4), CE score (0.0204), and VLM-as-Judge score (0.0).

178

179 **Case 2 Analysis.** Although the OCR output in Case 2 may appear superficially readable, it contains critical errors in financial and numerical
180 figures that cause high VLM disagreement. These character-level discrepancies—easily missed by semantic-only methods—result in elevated
181 CE values, correctly flagging the output as unreliable for downstream applications such as invoice processing or data extraction.



OCR Result (Qwen2.5-VL-72B OCR):

1. According to the first paragraph, the advantage of Antarctica is that it:
- \(\checkmark\)
- [A] A unique observation post for detecting important changes in the world's environment.
2. What would the result be if the West Antarctic Ice Sheet disappeared?
- \(\checkmark\)
- [B] The western part of the continent would be reduced.

Human Score:
4 (Mostly Incorrect)

VLM-as-Judge Score (Qwen2-VL-7B) ↑: 0.8917 ✗

CE Score (Qwen2.5-VL-72B+Qwen2-VL-7B) ↓: 0.9 ✓

Figure 9: Case 5: Multiple-choice test questions accurately evaluated by all methods, showing robust performance on educational content with specialized formatting.

These cases demonstrate that both CE and VLM-as-Judge methods generally align with human evaluation, but with important differences. CE shows particularly strong correlation with human judgments on complex documents (Cases 1, 3, and 5), where nuanced understanding is required. The strong agreement between human scores and automated methods validates our approach, particularly the training-free CE method which achieves comparable performance to supervised VLM-as-Judge approaches without requiring explicit quality assessment prompting.

Notably, in cases with clear quality issues (Case 2), both methods accurately identify poor OCR results, confirming their reliability for quality filtering applications. For structured content (Case 4), both approaches excel, suggesting that format preservation is well-captured by these evaluation techniques.

These qualitative examples complement our quantitative results in Section 4, reinforcing that CE provides a robust, unsupervised alternative to more complex and computationally expensive VLM-as-Judge methods for OCR quality assessment.

H Prompt

The effectiveness of OCR tasks in VLLMs is significantly influenced by the design of prompts. We present the primary prompts utilized in our experiments below, each serving a specific purpose in the evaluation framework. These prompts were carefully crafted based on extensive preliminary explorations to maximize OCR performance while maintaining consistency across different models.

OCR VLM-as-Judge Prompt

You are an expert evaluator assessing the quality of OCR (Optical Character Recognition) model predictions.

You will receive:

- A question
- A prediction generated by the OCR model.
- The corresponding image containing text.

Your task is to judge how well the predicted text matches the visual textual content in the image, with respect to the question's intent.

Evaluation criteria:

- (1) Focus only on whether the prediction correctly reflects the textual content of the image.
- (2) Assign a score from 0 to 1 in steps of 0.1, using the following four-level guideline:

Scoring Reference:

- **0.9-1.0 (Perfect Match)**
The prediction matches the image text exactly, with no errors or omissions.
- **0.7-0.8 (Minor Errors, Still Clear)**
The prediction is very close to the image text, with only small mistakes that do not affect understanding.
- **0.4-0.6 (Partially Correct)**
The prediction contains noticeable errors or captures only part of the text, reducing clarity.
- **0.0-0.3 (Mostly or Completely Incorrect)**
The prediction is largely incorrect or unrelated to the text in the image.

Respond only with the numerical score (e.g., 0.9). Do not include any explanation or commentary.

OCR CE-Guided Router Prompt

You are an expert AI assistant tasked with improving answers to visual questions. Please look at the image and examine the following question and the current answers from different models.
Question: {question}
Current model predictions: {predictions_str}
Your task is to synthesize these predictions and create a single improved answer that:

- (1) Is more accurate based on the visual content
- (2) Is concise and direct
- (3) Uses a natural, conversational tone
- (4) Maintains the core meaning of the original answers if they were correct
- (5) Improves clarity and precision

Do not invent details not present in the image. Your answer should be grounded in what is actually visible.
Please provide ONLY the improved answer with no explanations or additional text.

OCR Qwen2.5VL-72B Prompt

You are a very professional expert at OCR tasks. Please analyze the following image and extract all text content.

- (1) Ensure that the extracted text matches the original in the image and maintains the original structure.
- (2) If the image contains two columns, you should extract all text in the left column before you move on to the right.
- (3) Ignore the headers, but keep the footnotes, title.
- (4) You could either use LaTeX, KaTeX or Markdown format for math formulas, physics equations, chemical expressions etc.
- (5) Do not add or modify the original content!

H.1 Evaluation Metrics

To evaluate model performance, we employed multiple metrics assessing various aspects of OCR quality. The character error rate (CER) and word error rate (WER) measured the character-level and word-level accuracy respectively, capturing fine-grained recognition performance. BLEU and ROUGE-L scores assessed the overall textual similarity between predictions and ground truth, with ROUGE-L particularly sensitive to longer n-gram overlaps that indicate structural preservation. For semantic accuracy, we calculated embedding similarity using sentence transformers, which effectively captures meaning preservation even when exact wording differs. Additionally, we employed task-specific metrics for specialized OCR applications: formula recognition used a LaTeX-aware F1 score that accounts for equivalent expressions, while table extraction was evaluated using TEDS (Table Edit Distance Similarity) to measure structural fidelity. Collectively, these metrics provided a comprehensive evaluation framework that assessed both the syntactic accuracy and semantic fidelity of OCR outputs across diverse document understanding scenarios.

H.2 Details of Compared Methods

Our comparative analysis incorporated several state-of-the-art approaches for OCR evaluation and improvement. The VLM-as-Judge paradigm was implemented using three different foundation models: GPT4o, Qwen2-VL-72B, and Qwen2-VL-7B, each prompted with the standardized evaluation instructions shown in Section A.2. For traditional evaluation metrics, we utilized the benchmark methodology from OCRBench, employing exact match and fuzzy matching criteria with standardized preprocessing to normalize formatting variations. The self-verification methods included both the token-level confidence based approach and our proposed Consensus Entropy framework. For token-level confidence, we aggregated model-reported logit scores and calibrated them using temperature scaling. The CE framework was implemented with multiple variations in entropy calculation methods, including Mean Distance, Sum, Max, and Mean methods as detailed in Section 3.1 of the main paper. All baseline methods were evaluated using identical test samples and environmental configurations to ensure fair comparison.

I Ensemble Results

I.1 CCOCR

Table 11 summarizes the ensemble performance of three CE-selected model combinations on the CCOCR benchmark, spanning four major OCR task categories: *KIE*, *Document Parsing*, *Multilingual OCR*, and *Multi-scene OCR*. For each combination, we report both the ensemble results (first row) and the performance of the best-performing single model in the group (second row, *italicized*). The final column reports the *Overall Score*, computed as the average across the four tasks.

I.2 OCRBench-V2

Table 12 presents the bilingual evaluation of seven CE-selected multi-model aggregation schemes on OCRBench-V2. Each scheme combines four to five vision-language models (VLMs). *Ensemble Score* denotes the overall performance of the aggregated model, whereas *Single Best* is the highest-scoring individual model within the same ensemble for the corresponding language subset. Their absolute difference, $\Delta = \text{Ensemble Score} - \text{Single Best}$, is reported in the right-most column. Results with the highest score in each language subset are highlighted in blue, and all positive gains ($\Delta > 0$) are shaded in green.

Table 11: CCOCR Ensemble Results

Models	Task Performance				Overall
	KIE	Doc Parsing	Multi Language	Multi Scene	Score
Qwen2.5-VL-3B, Qwen2.5-VL-7B, Qwen2.5-VL-72B	89.27	64.85	79.75	85.56	79.86
Best Single (Qwen2.5-VL-72B)	89.51	62.34	80.77	86.34	79.74
Qwen2.5-VL-3B, Qwen2.5-VL-7B, Qwen2-VL-7B-Instruct	87.16	61.19	77.49	83.09	77.23
Best Single (Qwen2.5-VL-7B)	87.42	60.67	78.49	84.12	77.67
InternVL2_5-4B, Qwen2.5-VL-7B, Qwen2.5-VL-72B	88.71	62.48	79.90	84.90	79.00
Best Single (Qwen2.5-VL-72B)	89.51	62.34	80.77	86.34	79.74

Table 12: OCRBench-V2 Ensemble Results

Models	Ensemble Score		Single Best		Δ	
	English	Chinese	English	Chinese	English	Chinese
internvl2 5 26b, qwen2vl-8b, gemini pro MiniCPM-V-2 6	0.588	0.528	0.555	0.442	0.033	0.086
gemini pro, gpt4o, internvl2 5 26b qwen2vl-8b	0.589	0.521	0.555	0.442	0.034	0.079
internvl2 5 26b, gpt4o, qwen2vl-8b MiniCPM-V-2 6	0.567	0.483	0.555	0.442	0.012	0.041
internvl2_8b, cambrian_8b, llava_onevision_qwen2_7b_ov	0.538	0.395	0.404	0.363	0.134	0.032
idefics3, llava_onevision_qwen2_7b_ov, MiniCPM-V-2_6	0.550	0.466	0.416	0.307	0.134	0.159
internvl2_8b, cambrian_8b, llavar	0.521	0.387	0.404	0.363	0.117	0.024
internvl2_8b, cambrian_8b, textharmony	0.520	0.394	0.404	0.363	0.116	0.031

234 I.3 OCRBench

235 Tables 13, 14, 15 present supplementary results on OCRBENCH, covering CE-selected ensemble combinations of 3, 4, and 5 models. Each row
236 represents a specific ensemble configuration. The *Score* column reports the final ensemble performance obtained via CE-based prediction
237 selection. The *Max*, *Min*, and *Avg* columns correspond to the highest, lowest, and mean scores among the participating individual models.
238 The rightmost columns denote the absolute gains of the ensemble over its components: $\Delta_{\max} = \text{Score} - \text{Min}$, $\Delta_{\min} = \text{Score} - \text{Max}$, and
239 $\Delta_{\text{avg}} = \text{Score} - \text{Avg}$. Blue highlights indicate the best result among both the ensemble and its constituent models, while green shading marks
240 positive relative gains.

Table 13: OCRBench Ensemble Results (5 models)

Models	Score	Max	Min	Avg	Δ_{\max}	Δ_{\min}	Δ_{avg}
Ovis2-1B, Qwen2.5-VL-72B, SenseChat-Vision Step1V, Step1o	95.8	92.6	87.9	89.50	7.9	3.2	6.30
Ovis2-4B, Qwen2.5-VL-7B, SenseChat-Vision Step1V, Step1o	95.7	92.6	87.4	89.78	8.3	3.1	5.92
Ovis2-4B, Qwen2.5-VL-72B, SenseChat-Vision Step1V, Step1o	95.6	92.6	87.9	89.88	7.7	3.0	5.72
Ovis2-4B, Qwen-VL-Plus-0809, SenseChat-Vision							

continued on next page

continuation sheet 13:OCRBench Ensemble Results (5 models)

Models	Score	Max	Min	Avg	Δ_{max}	Δ_{min}	Δ_{avg}
Step1V, Step1o	95.6	92.6	84.3	89.16	11.3	3.0	6.44
MiniCPM-V-2_6, Qwen2.5-VL-72B, SenseChat-Vision Step1V, Step1o	95.6	92.6	85.2	88.74	10.4	3.0	6.86
Ovis2-1B, Qwen2.5-VL-7B, SenseChat-Vision Step1V, Step1o	95.5	92.6	87.4	89.40	8.1	2.9	6.10
Ovis2-4B, SenseChat-Vision, Step1V Step1o, bailingMM-Lite	95.5	92.6	85.2	89.34	10.3	2.9	6.16
GLM4V_PLUS, MUG-U-7B, Qwen2.5-VL-72B Step1V, Step1o	95.5	92.6	84.3	88.90	11.2	2.9	6.60
Ovis1.6-Gemma2-27B, Qwen2.5-VL-72B, SenseChat-Vision Step1V, Step1o	95.5	92.6	85.6	88.82	9.9	2.9	6.68
MUG-U-7B, Qwen2-VL-7B-Instruct, SenseChat-Vision Step1o, bailingMM-Lite	95.5	92.6	84.3	88.52	11.2	2.9	6.98
GLM4V_PLUS, Qwen2.5-VL-7B, SenseChat-Vision Step1V, Step1o	95.5	92.6	84.3	88.46	11.2	2.9	7.04
InternVL2_5-78B-MPO, Qwen2.5-VL-72B, SenseChat-Vision Step1V, Step1o	95.4	92.6	87.9	89.88	7.5	2.8	5.52
Ovis2-4B, Qwen2-VL-7B-Instruct, SenseChat-Vision Step1V, Step1o	95.4	92.6	84.3	89.16	11.1	2.8	6.24
Ovis2-4B, Qwen2-VL-7B-Instruct, Qwen2.5-VL-72B Step1V, Step1o	95.4	92.6	84.3	88.86	11.1	2.8	6.54
GLM4V_PLUS, Qwen2.5-VL-72B, SenseChat-Vision Step1V, Step1o	95.4	92.6	84.3	88.56	11.1	2.8	6.84
Ovis2-1B, Qwen2-VL-7B-Instruct, Qwen2.5-VL-72B Step1V, Step1o	95.4	92.6	84.3	88.48	11.1	2.8	6.92
MUG-U-7B, Ovis2-1B, Qwen2.5-VL-72B Step1V, Step1o	95.3	92.6	87.9	89.84	7.4	2.7	5.46
MUG-U-7B, Qwen2.5-VL-7B, SenseChat-Vision Step1V, Step1o	95.3	92.6	87.4	89.82	7.9	2.7	5.48
InternVL2_5-78B-MPO, Ovis2-1B, Qwen2.5-VL-72B Step1V, Step1o	95.3	92.6	87.9	89.80	7.4	2.7	5.50
MUG-U-7B, Ovis2-4B, Qwen-VL-Plus-0809 Step1V, Step1o	95.3	92.6	84.3	89.50	11.0	2.7	5.80
MUG-U-7B, Qwen2-VL-72B-Instruct, SenseChat-Vision Step1o, bailingMM-Lite	95.3	92.6	85.2	89.42	10.1	2.7	5.88
Ovis2-8B, SenseChat-Vision, Step1V Step1o, bailingMM-Lite	95.3	92.6	85.2	89.02	10.1	2.7	6.28
Ovis2-1B, Qwen-VL-Plus-0809, SenseChat-Vision Step1V, Step1o	95.3	92.6	84.3	88.78	11.0	2.7	6.52
Ovis2-1B, Qwen2-VL-7B-Instruct, SenseChat-Vision Step1V, Step1o	95.3	92.6	84.3	88.78	11.0	2.7	6.52
Qwen2.5-VL-72B, SenseChat-Vision, Step1V Step1o, bailingMM-Lite	95.3	92.6	85.2	88.74	10.1	2.7	6.56
MUG-U-7B, Qwen-VL-Plus-0809, SenseChat-Vision							

continued on next page

continuation sheet 13:OCRBench Ensemble Results (5 models)

Models	Score	Max	Min	Avg	Δ_{max}	Δ_{min}	Δ_{avg}
Step1o, bailingMM-Lite	95.3	92.6	84.3	88.52	11.0	2.7	6.78
GLM4V_PLUS, Ovis2-1B, Qwen2.5-VL-72B Step1V, Step1o	95.3	92.6	84.3	88.48	11.0	2.7	6.82
GLM4V_PLUS, MUG-U-7B, Qwen-VL-Plus-0809 Step1V, Step1o	95.3	92.6	84.3	88.18	11.0	2.7	7.12
MUG-U-7B, Ovis2-4B, Qwen2.5-VL-72B Step1V, Step1o	95.2	92.6	87.9	90.22	7.3	2.6	4.98
InternVL2_5-78B-MPO, Ovis2-1B, Qwen2.5-VL-7B Step1V, Step1o	95.2	92.6	87.4	89.70	7.8	2.6	5.50
MUG-U-7B, Qwen2-VL-72B-Instruct, Qwen2.5-VL-7B Step1V, Step1o	95.2	92.6	87.4	89.70	7.8	2.6	5.50
Ovis2-8B, Qwen2.5-VL-72B, SenseChat-Vision Step1V, Step1o	95.2	92.6	87.9	89.56	7.3	2.6	5.64
MUG-U-7B, Ovis2-4B, Qwen2-VL-7B-Instruct Step1V, Step1o	95.2	92.6	84.3	89.50	10.9	2.6	5.70
Ovis2-8B, Qwen2.5-VL-7B, SenseChat-Vision Step1V, Step1o	95.2	92.6	87.4	89.46	7.8	2.6	5.74
MUG-U-7B, SenseChat-Vision, Step1V Step1o, bailingMM-Lite	95.2	92.6	85.2	89.38	10.0	2.6	5.82
MUG-U-7B, Qwen2.5-VL-7B, SenseChat-Vision Step1o, bailingMM-Lite	95.2	92.6	85.2	89.14	10.0	2.6	6.06
MUG-U-7B, MiniCPM-V-2_6, Qwen2.5-VL-72B Step1V, Step1o	95.2	92.6	85.2	89.08	10.0	2.6	6.12
Ovis2-1B, SenseChat-Vision, Step1V Step1o, bailingMM-Lite	95.2	92.6	85.2	88.96	10.0	2.6	6.24
GLM4V_PLUS, MUG-U-7B, Qwen2.5-VL-7B Step1V, Step1o	95.2	92.6	84.3	88.80	10.9	2.6	6.40
Qwen2.5-VL-7B, SenseChat-Vision, Step1V Step1o, bailingMM-Lite	95.2	92.6	85.2	88.64	10.0	2.6	6.56
MiniCPM-V-2_6, Ovis2-1B, Qwen2.5-VL-7B Step1V, Step1o	95.2	92.6	85.2	88.56	10.0	2.6	6.64
Ovis2-1B, Qwen-VL-Plus-0809, Qwen2.5-VL-72B Step1V, Step1o	95.2	92.6	84.3	88.48	10.9	2.6	6.72
Qwen2-VL-7B-Instruct, SenseChat-Vision, Step1V Step1o, bailingMM-Lite	95.2	92.6	84.3	88.02	10.9	2.6	7.18
InternVL2_5-78B-MPO, MUG-U-7B, Qwen2.5-VL-72B Step1V, Step1o	95.1	92.6	87.9	90.22	7.2	2.5	4.88
MUG-U-7B, Ovis2-4B, Qwen2.5-VL-7B Step1V, Step1o	95.1	92.6	87.4	90.12	7.7	2.5	4.98
Ovis2-4B, Qwen2-VL-72B-Instruct, SenseChat-Vision Step1V, Step1o	95.1	92.6	88.6	90.06	6.5	2.5	5.04
InternVL2_5-78B-MPO, Qwen2.5-VL-7B, SenseChat-Vision Step1V, Step1o	95.1	92.6	87.4	89.78	7.7	2.5	5.32
MUG-U-7B, Ovis2-8B, Step1V							

continued on next page

continuation sheet 13:OCRBench Ensemble Results (5 models)

Models	Score	Max	Min	Avg	Δ_{max}	Δ_{min}	Δ_{avg}
Step1o, bailingMM-Lite	95.1	92.6	85.2	89.36	9.9	2.5	5.74
MUG-U-7B, Qwen-VL-Max-0809, SenseChat-Vision Step1o, bailingMM-Lite	95.1	92.6	85.2	89.28	9.9	2.5	5.82
MUG-U-7B, Qwen2.5-VL-72B, SenseChat-Vision Step1o, bailingMM-Lite	95.1	92.6	85.2	89.24	9.9	2.5	5.86
MUG-U-7B, Qwen2-VL-7B-Instruct, SenseChat-Vision Step1V, Step1o	95.1	92.6	84.3	89.20	10.8	2.5	5.90
Ovis2-4B, Ovis2-8B, Qwen2.5-VL-72B Step1o, bailingMM-Lite	95.1	92.6	85.2	89.18	9.9	2.5	5.92
MUG-U-7B, Ovis2-1B, Qwen-VL-Plus-0809 Step1V, Step1o	95.1	92.6	84.3	89.12	10.8	2.5	5.98
MUG-U-7B, Qwen-VL-Plus-0809, Qwen2-VL-72B-Instruct Step1V, Step1o	95.1	92.6	84.3	89.08	10.8	2.5	6.02
InternVL2_5-26B, Ovis2-1B, SenseChat-Vision Step1V, Step1o	95.1	92.6	85.4	89.00	9.7	2.5	6.10
Qwen2-VL-72B-Instruct, SenseChat-Vision, Step1V Step1o, bailingMM-Lite	95.1	92.6	85.2	88.92	9.9	2.5	6.18
InternVL2_5-38B, MUG-U-7B, Qwen2.5-VL-7B SenseChat-Vision, Step1o	95.1	92.6	84.1	88.92	11.0	2.5	6.18
Ovis2-4B, Qwen-VL-Plus-0809, Qwen2.5-VL-72B Step1V, Step1o	95.1	92.6	84.3	88.86	10.8	2.5	6.24
Ovis2-8B, Qwen-VL-Plus-0809, SenseChat-Vision Step1V, Step1o	95.1	92.6	84.3	88.84	10.8	2.5	6.26
Ovis2-8B, Qwen2-VL-7B-Instruct, SenseChat-Vision Step1V, Step1o	95.1	92.6	84.3	88.84	10.8	2.5	6.26
MUG-U-7B, Qwen2-VL-7B-Instruct, Qwen2.5-VL-7B Step1V, Step1o	95.1	92.6	84.3	88.80	10.8	2.5	6.30
InternVL2_5-38B, MUG-U-7B, Qwen2.5-VL-7B Step1V, Step1o	95.1	92.6	84.1	88.76	11.0	2.5	6.34
MiniCPM-V-2_6, Ovis2-1B, Qwen2.5-VL-72B Step1V, Step1o	95.1	92.6	85.2	88.66	9.9	2.5	6.44
Ovis2-1B, Qwen2.5-VL-72B, Step1V Step1o, bailingMM-Lite	95.1	92.6	85.2	88.66	9.9	2.5	6.44
MiniCPM-V-2_6, Qwen2.5-VL-7B, SenseChat-Vision Step1V, Step1o	95.1	92.6	85.2	88.64	9.9	2.5	6.46
Ovis2-1B, Qwen2.5-VL-7B, Step1V Step1o, bailingMM-Lite	95.1	92.6	85.2	88.56	9.9	2.5	6.54
Qwen-VL-Plus-0809, Qwen2.5-VL-7B, SenseChat-Vision Step1V, Step1o	95.1	92.6	84.3	88.46	10.8	2.5	6.64
GLM4V_PLUS, Ovis2-1B, Qwen2.5-VL-7B Step1V, Step1o	95.1	92.6	84.3	88.38	10.8	2.5	6.72
MUG-U-7B, Qwen-VL-Plus-0809, Step1V Step1o, bailingMM-Lite	95.1	92.6	84.3	88.36	10.8	2.5	6.74
GLM4V_PLUS, Ovis2-4B, Qwen-VL-Plus-0809							

continued on next page

continuation sheet 13:OCRBench Ensemble Results (5 models)

Models	Score	Max	Min	Avg	Δ_{max}	Δ_{min}	Δ_{avg}
SenseChat-Vision, Step1o	95.1	92.6	84.3	88.30	10.8	2.5	6.80
GLM4V_PLUS, SenseChat-Vision, Step1V Step1o, bailingMM-Lite	95.1	92.6	84.3	88.02	10.8	2.5	7.08
GLM4V_PLUS, Qwen2-VL-72B-Instruct, Step1V Step1o, bailingMM-Lite	95.1	92.6	84.3	87.90	10.8	2.5	7.20
MUG-U-7B, Qwen2.5-VL-72B, SenseChat-Vision Step1V, Step1o	95.0	92.6	87.9	89.92	7.1	2.4	5.08
MUG-U-7B, Ovis2-8B, Qwen2.5-VL-72B Step1V, Step1o	95.0	92.6	87.9	89.90	7.1	2.4	5.10
Ovis2-4B, Qwen2-VL-72B-Instruct, Qwen2.5-VL-72B Step1V, Step1o	95.0	92.6	87.9	89.76	7.1	2.4	5.24
MUG-U-7B, Ovis2-1B, Qwen2.5-VL-7B Step1V, Step1o	95.0	92.6	87.4	89.74	7.6	2.4	5.26
MUG-U-7B, Ovis2-4B, Step1V Step1o, bailingMM-Lite	95.0	92.6	85.2	89.68	9.8	2.4	5.32
InternVL2_5-38B, MUG-U-7B, Ovis2-4B Step1V, Step1o	95.0	92.6	84.1	89.46	10.9	2.4	5.54
MUG-U-7B, Ovis2-1B, Step1V Step1o, bailingMM-Lite	95.0	92.6	85.2	89.30	9.8	2.4	5.70
Ovis2-34B, Ovis2-4B, Qwen2.5-VL-72B SenseChat-Vision, Step1V	95.0	90.9	87.9	89.30	7.1	4.1	5.70
MUG-U-7B, Qwen2-VL-72B-Instruct, Step1V Step1o, bailingMM-Lite	95.0	92.6	85.2	89.26	9.8	2.4	5.74
InternVL2_5-78B-MPO, Ovis2-1B, Qwen2-VL-7B-Instruct SenseChat-Vision, Step1o	95.0	92.6	84.3	89.24	10.7	2.4	5.76
Ovis2-4B, Qwen2.5-VL-72B, SenseChat-Vision Step1o, bailingMM-Lite	95.0	92.6	85.2	89.20	9.8	2.4	5.80
MUG-U-7B, Ovis2-8B, Qwen2-VL-7B-Instruct Step1V, Step1o	95.0	92.6	84.3	89.18	10.7	2.4	5.82
InternVL2_5-78B-MPO, Qwen-VL-Plus-0809, SenseChat-Vision Step1V, Step1o	95.0	92.6	84.3	89.16	10.7	2.4	5.84
Ovis2-4B, Qwen2.5-VL-72B, Step1V Step1o, bailingMM-Lite	95.0	92.6	85.2	89.04	9.8	2.4	5.96
InternVL2_5-38B, MUG-U-7B, Qwen2-VL-72B-Instruct Step1V, Step1o	95.0	92.6	84.1	89.04	10.9	2.4	5.96
MUG-U-7B, NVLM, Qwen2.5-VL-72B Step1V, Step1o	95.0	92.6	84.9	89.02	10.1	2.4	5.98
MiniCPM-V-2_6, Ovis2-1B, Qwen2.5-VL-72B SenseChat-Vision, Step1o	95.0	92.6	85.2	88.82	9.8	2.4	6.18
InternVL2_5-78B, Qwen2.5-VL-7B, SenseChat-Vision Step1V, Step1o	95.0	92.6	85.3	88.66	9.7	2.4	6.34
Ovis2-8B, Qwen2-VL-7B-Instruct, Qwen2.5-VL-72B Step1V, Step1o	95.0	92.6	84.3	88.54	10.7	2.4	6.46
InternVL2_5-78B, Ovis2-4B, Qwen-VL-Plus-0809							

continued on next page

continuation sheet 13: OCRBench Ensemble Results (5 models)

Models	Score	Max	Min	Avg	Δ_{max}	Δ_{min}	Δ_{avg}
SenseChat-Vision, Step1o	95.0	92.6	84.3	88.50	10.7	2.4	6.50
InternVL2_5-38B, Ovis2-1B, Qwen2.5-VL-7B SenseChat-Vision, Step1o	95.0	92.6	84.1	88.50	10.9	2.4	6.50
Ovis2-4B, Qwen-VL-Plus-0809, Step1V Step1o, bailingMM-Lite	95.0	92.6	84.3	88.32	10.7	2.4	6.68
GLM4V_PLUS, MUG-U-7B, Qwen2-VL-7B-Instruct Step1V, Step1o	95.0	92.6	84.3	88.18	10.7	2.4	6.82
MiniCPM-V-2_6, Ovis2-1B, Qwen2-VL-7B-Instruct SenseChat-Vision, Step1o	95.0	92.6	84.3	88.10	10.7	2.4	6.90
MiniCPM-V-2_6, Qwen-VL-Plus-0809, SenseChat-Vision Step1V, Step1o	95.0	92.6	84.3	88.02	10.7	2.4	6.98
MiniCPM-V-2_6, Ovis2-1B, Qwen-VL-Plus-0809 Step1V, Step1o	95.0	92.6	84.3	87.94	10.7	2.4	7.06
Ovis2-1B, Qwen-VL-Plus-0809, Step1V Step1o, bailingMM-Lite	95.0	92.6	84.3	87.94	10.7	2.4	7.06
GLM4V_PLUS, Qwen-VL-Plus-0809, SenseChat-Vision Step1V, Step1o	95.0	92.6	84.3	87.84	10.7	2.4	7.16

Table 14: OCRBench Ensemble Results (4 models)

Models	Score	Max	Min	Avg	Δ_{max}	Δ_{min}	Δ_{avg}
Ovis2-1B, Qwen2.5-VL-7B, Step1V Step1o	95.5	92.6	87.4	89.40	8.1	2.9	6.10
Qwen2.5-VL-72B, SenseChat-Vision, Step1V Step1o	95.4	92.6	87.9	89.62	7.5	2.8	5.78
MUG-U-7B, Qwen2.5-VL-72B, Step1V Step1o	95.3	92.6	87.9	90.05	7.4	2.7	5.25
MUG-U-7B, Qwen-VL-Plus-0809, Step1V Step1o	95.1	92.6	84.3	89.15	10.8	2.5	5.95
Ovis2-1B, Qwen-VL-Plus-0809, Step1V Step1o	95.1	92.6	84.3	88.62	10.8	2.5	6.47
Ovis2-1B, Qwen2-VL-72B-Instruct, SenseChat-Vision Step1o	95.0	92.6	88.8	89.95	6.2	2.4	5.05
Ovis2-1B, Qwen2.5-VL-72B, Step1V Step1o	95.0	92.6	87.9	89.53	7.1	2.4	5.47
Ovis2-1B, Qwen2-VL-7B-Instruct, SenseChat-Vision Step1o	95.0	92.6	84.3	88.83	10.7	2.4	6.17
MUG-U-7B, Qwen2.5-VL-7B, Step1V Step1o	94.9	92.6	87.4	89.92	7.5	2.3	4.97
Ovis2-1B, SenseChat-Vision, Step1V Step1o	94.9	92.6	88.6	89.90	6.3	2.3	5.00
Ovis2-1B, Qwen2.5-VL-72B, SenseChat-Vision							

continued on next page

continuation sheet 14: OCRBench Ensemble Results (4 models)

Models	Score	Max	Min	Avg	Δ_{max}	Δ_{min}	Δ_{avg}
Step1o	94.9	92.6	87.9	89.72	7.0	2.3	5.17
Qwen2.5-VL-7B, SenseChat-Vision, Step1V Step1o	94.9	92.6	87.4	89.50	7.5	2.3	5.40
SenseChat-Vision, Step1V, Step1o bailingMM-Lite	94.9	92.6	85.2	88.95	9.7	2.3	5.95
Ovis2-1B, Qwen2-VL-7B-Instruct, Step1V Step1o	94.9	92.6	84.3	88.62	10.6	2.3	6.28
MUG-U-7B, Ovis2-4B, Step1V Step1o	94.8	92.6	88.6	90.80	6.2	2.2	4.00
MUG-U-7B, Qwen2-VL-72B-Instruct, Step1V Step1o	94.8	92.6	88.6	90.28	6.2	2.2	4.53
Ovis2-4B, Qwen2.5-VL-7B, Step1V Step1o	94.8	92.6	87.4	89.88	7.4	2.2	4.92
Ovis2-1B, Qwen-VL-Max-0809, SenseChat-Vision Step1o	94.8	92.6	88.1	89.78	6.7	2.2	5.03
MUG-U-7B, SenseChat-Vision, Step1o bailingMM-Lite	94.8	92.6	85.2	89.58	9.6	2.2	5.22
Ovis2-4B, Qwen-VL-Plus-0809, SenseChat-Vision Step1o	94.8	92.6	84.3	89.30	10.5	2.2	5.50
MUG-U-7B, Qwen2-VL-7B-Instruct, Step1V Step1o	94.8	92.6	84.3	89.15	10.5	2.2	5.65
Qwen2-VL-72B-Instruct, SenseChat-Vision, Step1o bailingMM-Lite	94.8	92.6	85.2	89.00	9.6	2.2	5.80
MUG-U-7B, Ovis2-8B, Step1V Step1o	94.7	92.6	88.6	90.40	6.1	2.1	4.30
MUG-U-7B, Qwen2.5-VL-72B, SenseChat-Vision Step1o	94.7	92.6	87.9	90.25	6.8	2.1	4.45
Qwen2-VL-72B-Instruct, SenseChat-Vision, Step1V Step1o	94.7	92.6	88.6	89.85	6.1	2.1	4.85
MUG-U-7B, Qwen2.5-VL-72B, SenseChat-Vision Step1V	94.7	91.1	87.9	89.25	6.8	3.6	5.45
InternVL2_5-26B, Qwen2-VL-72B-Instruct, Step1V Step1o	94.7	92.6	85.4	88.85	9.3	2.1	5.85
Ovis2-1B, Qwen-VL-Plus-0809, SenseChat-Vision Step1o	94.7	92.6	84.3	88.83	10.4	2.1	5.88
Ovis2-1B, Qwen2.5-VL-72B, SenseChat-Vision Step1V	94.7	89.4	87.9	88.72	6.8	5.3	5.97
Qwen2-VL-7B-Instruct, SenseChat-Vision, Step1o bailingMM-Lite	94.7	92.6	84.3	87.88	10.4	2.1	6.83
MUG-U-7B, Ovis2-1B, Step1V Step1o	94.6	92.6	88.6	90.33	6.0	2.0	4.28
MUG-U-7B, Ovis2-1B, Qwen2.5-VL-72B Step1o	94.6	92.6	87.9	90.15	6.7	2.0	4.45
MUG-U-7B, Qwen2.5-VL-7B, SenseChat-Vision							

continued on next page

continuation sheet 14: OCRBench Ensemble Results (4 models)

Models	Score	Max	Min	Avg	Δ_{max}	Δ_{min}	Δ_{avg}
Step1o	94.6	92.6	87.4	90.12	7.2	2.0	4.47
Ovis2-4B, Qwen2.5-VL-72B, Step1V Step1o	94.6	92.6	87.9	90.00	6.7	2.0	4.60
Ovis2-1B, Qwen2-VL-72B-Instruct, Step1V Step1o	94.6	92.6	88.6	89.75	6.0	2.0	4.85
Qwen-VL-Max-0809, SenseChat-Vision, Step1V Step1o	94.6	92.6	88.1	89.67	6.5	2.0	4.92
Ovis2-1B, Qwen2.5-VL-7B, SenseChat-Vision Step1o	94.6	92.6	87.4	89.60	7.2	2.0	5.00
Ovis2-8B, Qwen2.5-VL-7B, Step1V Step1o	94.6	92.6	87.4	89.47	7.2	2.0	5.12
Ovis2-4B, Qwen-VL-Plus-0809, Step1V Step1o	94.6	92.6	84.3	89.10	10.3	2.0	5.50
Qwen-VL-Max-0809, SenseChat-Vision, Step1o bailingMM-Lite	94.6	92.6	85.2	88.83	9.4	2.0	5.78
Qwen2-VL-72B-Instruct, Step1V, Step1o bailingMM-Lite	94.6	92.6	85.2	88.80	9.4	2.0	5.80
Qwen-VL-Plus-0809, SenseChat-Vision, Step1V Step1o	94.6	92.6	84.3	88.72	10.3	2.0	5.88
MUG-U-7B, Ovis2-4B, Qwen2.5-VL-72B Step1o	94.5	92.6	87.9	90.62	6.6	1.9	3.88
MUG-U-7B, Qwen-VL-Max-0809, Step1V Step1o	94.5	92.6	88.1	90.10	6.4	1.9	4.40
Ovis2-34B, Qwen2.5-VL-7B, Step1V Step1o	94.5	92.6	87.4	89.58	7.1	1.9	4.92
MUG-U-7B, Qwen2.5-VL-72B, Step1o bailingMM-Lite	94.5	92.6	85.2	89.20	9.3	1.9	5.30
Ovis2-4B, Qwen2.5-VL-72B, SenseChat-Vision Step1V	94.5	90.9	87.9	89.20	6.6	3.6	5.30
InternVL2_5-38B, MUG-U-7B, Step1V Step1o	94.5	92.6	84.1	89.10	10.4	1.9	5.40
Qwen2.5-VL-72B, SenseChat-Vision, Step1o bailingMM-Lite	94.5	92.6	85.2	88.78	9.3	1.9	5.72
Qwen2-VL-7B-Instruct, SenseChat-Vision, Step1V Step1o	94.5	92.6	84.3	88.72	10.2	1.9	5.78
GLM4V_PLUS, Ovis2-1B, Step1V Step1o	94.5	92.6	84.3	88.62	10.2	1.9	5.88
Ovis2-1B, Qwen2.5-VL-7B, SenseChat-Vision Step1V	94.5	89.4	87.4	88.60	7.1	5.1	5.90
InternVL2_5-38B, Ovis2-1B, Step1V Step1o	94.5	92.6	84.1	88.58	10.4	1.9	5.92
InternVL2_5-38B, Qwen2.5-VL-7B, Step1V Step1o	94.5	92.6	84.1	88.17	10.4	1.9	6.33
Ovis2-1B, Qwen-VL-Plus-0809, SenseChat-Vision							

continued on next page

continuation sheet 14: OCRBench Ensemble Results (4 models)

Models	Score	Max	Min	Avg	Δ_{max}	Δ_{min}	Δ_{avg}
Step1V	94.5	89.4	84.3	87.83	10.2	5.1	6.67
InternVL2_5-78B-MPO, Ovis2-8B, SenseChat-Vision Step1o	94.4	92.6	89.3	90.55	5.1	1.8	3.85
MUG-U-7B, Qwen2-VL-72B-Instruct, SenseChat-Vision Step1o	94.4	92.6	88.8	90.47	5.6	1.8	3.92
Ovis2-4B, Qwen2-VL-72B-Instruct, SenseChat-Vision Step1o	94.4	92.6	88.8	90.42	5.6	1.8	3.98
MUG-U-7B, Ovis2-1B, Qwen2.5-VL-7B Step1o	94.4	92.6	87.4	90.03	7.0	1.8	4.38
MUG-U-7B, Ovis2-4B, Qwen2-VL-72B-Instruct Qwen2.5-VL-72B	94.4	91.1	87.9	89.67	6.5	3.3	4.72
MUG-U-7B, Ovis2-4B, Qwen2.5-VL-72B Step1V	94.4	91.1	87.9	89.62	6.5	3.3	4.78
Ovis2-1B, Qwen-VL-Max-0809, Step1V Step1o	94.4	92.6	88.1	89.58	6.3	1.8	4.83
MUG-U-7B, Ovis2-1B, Qwen2-VL-72B-Instruct Step1V	94.4	91.1	88.6	89.38	5.8	3.3	5.03
MUG-U-7B, Qwen-VL-Plus-0809, SenseChat-Vision Step1o	94.4	92.6	84.3	89.35	10.1	1.8	5.05
MUG-U-7B, Qwen2-VL-7B-Instruct, SenseChat-Vision Step1o	94.4	92.6	84.3	89.35	10.1	1.8	5.05
Ovis2-4B, Qwen2-VL-7B-Instruct, SenseChat-Vision Step1o	94.4	92.6	84.3	89.30	10.1	1.8	5.10
MiniCPM-V-2_6, Ovis2-8B, SenseChat-Vision Step1o	94.4	92.6	85.2	89.12	9.2	1.8	5.28
Qwen-VL-Max-0809, Step1V, Step1o bailingMM-Lite	94.4	92.6	85.2	88.62	9.2	1.8	5.78
InternVL2-76B, Ovis2-1B, Step1V Step1o	94.4	92.6	84.2	88.60	10.2	1.8	5.80
MiniCPM-V-2_6, Qwen2.5-VL-7B, Step1V Step1o	94.4	92.6	85.2	88.45	9.2	1.8	5.95
Qwen2.5-VL-7B, Step1V, Step1o bailingMM-Lite	94.4	92.6	85.2	88.45	9.2	1.8	5.95
GLM4V_PLUS, Qwen2.5-VL-7B, Step1V Step1o	94.4	92.6	84.3	88.22	10.1	1.8	6.17
InternVL2_5-78B-MPO, Ovis2-4B, Step1V Step1o	94.3	92.6	88.6	90.75	5.7	1.7	3.55
Ovis2-4B, Qwen-VL-Max-0809, SenseChat-Vision Step1o	94.3	92.6	88.1	90.25	6.2	1.7	4.05
MUG-U-7B, Ovis2-4B, Step1o bailingMM-Lite	94.3	92.6	85.2	89.95	9.1	1.7	4.35
InternVL2_5-78B-MPO, Qwen2.5-VL-7B, Step1V Step1o	94.3	92.6	87.4	89.88	6.9	1.7	4.42
Ovis2-8B, Qwen2.5-VL-72B, Step1V							

continued on next page

continuation sheet 14: OCRBench Ensemble Results (4 models)

Models	Score	Max	Min	Avg	Δ_{max}	Δ_{min}	Δ_{avg}
Step1o	94.3	92.6	87.9	89.60	6.4	1.7	4.70
Ovis2-4B, Qwen2-VL-72B-Instruct, Step1o bailingMM-Lite	94.3	92.6	85.2	89.38	9.1	1.7	4.92
MiniCPM-V-2_6, Ovis2-4B, Step1V Step1o	94.3	92.6	85.2	89.33	9.1	1.7	4.97
MUG-U-7B, Ovis1.6-Gemma2-27B, Qwen2.5-VL-72B Step1o	94.3	92.6	85.6	89.30	8.7	1.7	5.00
Ovis2-4B, Qwen2.5-VL-72B, Step1o bailingMM-Lite	94.3	92.6	85.2	89.15	9.1	1.7	5.15
MUG-U-7B, Qwen2.5-VL-7B, SenseChat-Vision Step1V	94.3	91.1	87.4	89.12	6.9	3.2	5.17
MUG-U-7B, Qwen2.5-VL-7B, Step1o bailingMM-Lite	94.3	92.6	85.2	89.08	9.1	1.7	5.22
Ovis2-1B, Qwen2-VL-72B-Instruct, Step1o bailingMM-Lite	94.3	92.6	85.2	88.90	9.1	1.7	5.40
Ovis2-1B, Step1V, Step1o bailingMM-Lite	94.3	92.6	85.2	88.85	9.1	1.7	5.45
GLM4V_PLUS, SenseChat-Vision, Step1V Step1o	94.3	92.6	84.3	88.72	10.0	1.7	5.58
InternVL2_5-26B, Qwen-VL-Max-0809, Step1V Step1o	94.3	92.6	85.4	88.67	8.9	1.7	5.62
InternVL2_5-26B, Qwen2.5-VL-72B, Step1V Step1o	94.3	92.6	85.4	88.62	8.9	1.7	5.67
Qwen2.5-VL-72B, Step1V, Step1o bailingMM-Lite	94.3	92.6	85.2	88.58	9.1	1.7	5.72
GLM4V_PLUS, Qwen2-VL-72B-Instruct, Step1V Step1o	94.3	92.6	84.3	88.58	10.0	1.7	5.72
Ovis2-1B, Qwen2.5-VL-7B, Step1o bailingMM-Lite	94.3	92.6	85.2	88.55	9.1	1.7	5.75
MUG-U-7B, Ovis2-4B, Qwen2.5-VL-7B Step1o	94.2	92.6	87.4	90.50	6.8	1.6	3.70
InternVL2_5-78B-MPO, Qwen2-VL-72B-Instruct, SenseChat-Vision Step1o	94.2	92.6	88.8	90.42	5.4	1.6	3.77
Ovis2-4B, Ovis2-8B, Step1V Step1o	94.2	92.6	88.6	90.35	5.6	1.6	3.85
InternVL2_5-78B-MPO, Qwen2-VL-72B-Instruct, Step1V Step1o	94.2	92.6	88.6	90.22	5.6	1.6	3.98
Ovis2-4B, Qwen2-VL-72B-Instruct, Step1V Step1o	94.2	92.6	88.6	90.22	5.6	1.6	3.98
Ovis2-4B, Qwen2.5-VL-72B, SenseChat-Vision Step1o	94.2	92.6	87.9	90.20	6.3	1.6	4.00
InternVL2_5-78B-MPO, MiniCPM-V-2_6, Ovis2-4B Step1o	94.2	92.6	85.2	89.90	9.0	1.6	4.30
InternVL2_5-78B-MPO, MUG-U-7B, Qwen-VL-Plus-0809							

continued on next page

continuation sheet 14: OCRBench Ensemble Results (4 models)

Models	Score	Max	Min	Avg	Δ_{max}	Δ_{min}	Δ_{avg}
Step1o	94.2	92.6	84.3	89.72	9.9	1.6	4.47
InternVL2_5-78B-MPO, Ovis2-4B, Qwen2-VL-7B-Instruct Step1o	94.2	92.6	84.3	89.67	9.9	1.6	4.53

Table 15: OCRBench Ensemble Results (3 models)

Models	Score	Max	Min	Avg	Δ_{max}	Δ_{min}	Δ_{avg}
Ovis2-8B, SenseChat-Vision, Step1o	94.1	92.6	89.3	90.43	4.8	1.5	3.67
Qwen2.5-VL-7B, Step1V, Step1o	94.0	92.6	87.4	89.53	6.6	1.4	4.47
Ovis2-1B, Step1V, Step1o	94.0	92.6	88.6	90.07	5.4	1.4	3.93
Qwen2-VL-72B-Instruct, Step1V, Step1o	94.0	92.6	88.6	90.00	5.4	1.4	4.00
Ovis2-4B, Qwen2-VL-72B-Instruct, Step1o	94.0	92.6	88.8	90.77	5.2	1.4	3.23
MUG-U-7B, Step1V, Step1o	93.9	92.6	88.6	90.77	5.3	1.3	3.13
MUG-U-7B, Ovis2-4B, Step1o	93.9	92.6	90.9	91.53	3.0	1.3	2.37
MUG-U-7B, Qwen-VL-Plus-0809, Step1o	93.8	92.6	84.3	89.33	9.5	1.2	4.47
Ovis2-4B, Qwen2.5-VL-7B, Step1o	93.8	92.6	87.4	90.30	6.4	1.2	3.50
SenseChat-Vision, Step1V, Step1o	93.8	92.6	88.6	90.20	5.2	1.2	3.60
Ovis2-4B, Qwen-VL-Plus-0809, Step1o	93.7	92.6	84.3	89.27	9.4	1.1	4.43
Ovis2-4B, Step1V, Step1o	93.7	92.6	88.6	90.70	5.1	1.1	3.00
Qwen2-VL-72B-Instruct, SenseChat-Vision, Step1o	93.7	92.6	88.8	90.27	4.9	1.1	3.43
MUG-U-7B, SenseChat-Vision, Step1o	93.7	92.6	89.4	91.03	4.3	1.1	2.67
Ovis2-4B, Qwen2-VL-7B-Instruct, Step1o	93.6	92.6	84.3	89.27	9.3	1.0	4.33
Qwen-VL-Plus-0809, Step1V, Step1o	93.6	92.6	84.3	88.50	9.3	1.0	5.10
Ovis2-4B, Step1o, bailingMM-Lite	93.6	92.6	85.2	89.57	8.4	1.0	4.03
SenseChat-Vision, Step1o, bailingMM-Lite	93.6	92.6	85.2	89.07	8.4	1.0	4.53
Qwen2-VL-72B-Instruct, Step1o, bailingMM-Lite	93.6	92.6	85.2	88.87	8.4	1.0	4.73
MUG-U-7B, SenseChat-Vision, bailingMM-Lite	93.6	91.1	85.2	88.57	8.4	2.5	5.03
MUG-U-7B, Ovis2-4B, Qwen2.5-VL-72B	93.6	91.1	87.9	89.97	5.7	2.5	3.63
Ovis2-4B, Qwen-VL-Max-0809, Step1o	93.6	92.6	88.1	90.53	5.5	1.0	3.07
Qwen-VL-Max-0809, Step1V, Step1o	93.6	92.6	88.1	89.77	5.5	1.0	3.83
Ovis2-8B, Step1V, Step1o	93.6	92.6	88.6	90.17	5.0	1.0	3.43
Ovis2-4B, SenseChat-Vision, Step1o	93.6	92.6	89.4	90.97	4.2	1.0	2.63
Qwen-VL-Plus-0809, SenseChat-Vision, Step1o	93.5	92.6	84.3	88.77	9.2	0.9	4.73
Qwen2-VL-7B-Instruct, Step1V, Step1o	93.5	92.6	84.3	88.50	9.2	0.9	5.00
MUG-U-7B, Ovis1.6-Gemma2-27B, Step1o	93.5	92.6	85.6	89.77	7.9	0.9	3.73
MUG-U-7B, Qwen2.5-VL-7B, Step1o	93.5	92.6	87.4	90.37	6.1	0.9	3.13
Qwen2.5-VL-72B, SenseChat-Vision, Step1o	93.5	92.6	87.9	89.97	5.6	0.9	3.53

continued on next page

continuation sheet 15: OCRBench Ensemble Results (3 models)

Models	Score	Max	Min	Avg	Δ_{max}	Δ_{min}	Δ_{avg}
MUG-U-7B, Ovis2-4B, Step1V	93.5	91.1	88.6	90.20	4.9	2.4	3.30
MUG-U-7B, Step1o, bailingMM-Lite	93.4	92.6	85.2	89.63	8.2	0.8	3.77
Ovis2-34B, Step1V, Step1o	93.4	92.6	88.6	90.30	4.8	0.8	3.10
MUG-U-7B, Qwen2-VL-72B-Instruct, Step1V	93.4	91.1	88.6	89.50	4.8	2.3	3.90
MUG-U-7B, Qwen2-VL-72B-Instruct, Step1o	93.4	92.6	88.8	90.83	4.6	0.8	2.57
MUG-U-7B, Qwen2-VL-7B-Instruct, Step1o	93.3	92.6	84.3	89.33	9.0	0.7	3.97
Ovis2-8B, Step1o, bailingMM-Lite	93.3	92.6	85.2	89.03	8.1	0.7	4.27
Step1V, Step1o, bailingMM-Lite	93.3	92.6	85.2	88.80	8.1	0.7	4.50
Qwen2.5-VL-72B, Step1o, bailingMM-Lite	93.3	92.6	85.2	88.57	8.1	0.7	4.73
Ovis2-8B, Qwen2.5-VL-7B, Step1o	93.3	92.6	87.4	89.77	5.9	0.7	3.53
MUG-U-7B, Qwen2.5-VL-7B, Step1V	93.3	91.1	87.4	89.03	5.9	2.2	4.27
Ovis2-4B, Qwen2.5-VL-72B, Step1o	93.3	92.6	87.9	90.47	5.4	0.7	2.83
Ovis2-8B, Qwen2-VL-72B-Instruct, Step1o	93.3	92.6	88.8	90.23	4.5	0.7	3.07
Ovis2-1B, SenseChat-Vision, Step1o	93.3	92.6	89.0	90.33	4.3	0.7	2.97
Ovis2-34B, SenseChat-Vision, Step1o	93.3	92.6	89.4	90.57	3.9	0.7	2.73
InternVL2_5-38B, Step1V, Step1o	93.2	92.6	84.1	88.43	9.1	0.6	4.77
NVLM, Ovis2-4B, Step1o	93.2	92.6	84.9	89.47	8.3	0.6	3.73
Qwen2.5-VL-7B, SenseChat-Vision, Step1o	93.2	92.6	87.4	89.80	5.8	0.6	3.40
Qwen2.5-VL-72B, Step1V, Step1o	93.2	92.6	87.9	89.70	5.3	0.6	3.50
Qwen-VL-Max-0809, SenseChat-Vision, Step1o	93.2	92.6	88.1	90.03	5.1	0.6	3.17
MUG-U-7B, Ovis2-4B, Qwen2-VL-72B-Instruct	93.2	91.1	88.8	90.27	4.4	2.1	2.93
Ovis2-1B, Qwen2-VL-72B-Instruct, Step1o	93.2	92.6	88.8	90.13	4.4	0.6	3.07
MUG-U-7B, Ovis2-8B, Step1o	93.2	92.6	89.3	91.00	3.9	0.6	2.20
MUG-U-7B, NVLM, Step1o	93.1	92.6	84.9	89.53	8.2	0.5	3.57
Qwen-VL-Max-0809, Step1o, bailingMM-Lite	93.1	92.6	85.2	88.63	7.9	0.5	4.47
Qwen2.5-VL-7B, Step1o, bailingMM-Lite	93.1	92.6	85.2	88.40	7.9	0.5	4.70
MUG-U-7B, Ovis2-4B, Qwen2.5-VL-7B	93.1	91.1	87.4	89.80	5.7	2.0	3.30
MUG-U-7B, Qwen2.5-VL-72B, Step1o	93.1	92.6	87.9	90.53	5.2	0.5	2.57
MUG-U-7B, Qwen-VL-Max-0809, Step1o	93.1	92.6	88.1	90.60	5.0	0.5	2.50
Ovis2-8B, Qwen-VL-Max-0809, Step1o	93.1	92.6	88.1	90.00	5.0	0.5	3.10
MUG-U-7B, SenseChat-Vision, Step1V	93.1	91.1	88.6	89.70	4.5	2.0	3.40
GLM4V_PLUS, MUG-U-7B, Step1o	93.0	92.6	84.3	89.33	8.7	0.4	3.67
MUG-U-7B, Qwen-VL-Plus-0809, Step1V	93.0	91.1	84.3	88.00	8.7	1.9	5.00
NVLM, Step1V, Step1o	93.0	92.6	84.9	88.70	8.1	0.4	4.30
MUG-U-7B, Ovis2-4B, Qwen-VL-Max-0809	93.0	91.1	88.1	90.03	4.9	1.9	2.97
MUG-U-7B, Qwen-VL-Max-0809, Step1V	93.0	91.1	88.1	89.27	4.9	1.9	3.73
MUG-U-7B, Ovis2-34B, Step1o	93.0	92.6	89.7	91.13	3.3	0.4	1.87

continued on next page

continuation sheet 15: OCRBench Ensemble Results (3 models)

Models	Score	Max	Min	Avg	Δ_{max}	Δ_{min}	Δ_{avg}
InternVL2_5-78B-MPO, MUG-U-7B, Step1o	93.0	92.6	90.9	91.53	2.1	0.4	1.47
Qwen-VL-Plus-0809, Step1o, bailingMM-Lite	92.9	92.6	84.3	87.37	8.6	0.3	5.53
MUG-U-7B, Ovis2-1B, bailingMM-Lite	92.9	91.1	85.2	88.43	7.7	1.8	4.47
InternVL2_5-26B, Ovis2-4B, Step1o	92.9	92.6	85.4	89.63	7.5	0.3	3.27
MUG-U-7B, Qwen2.5-VL-72B, SenseChat-Vision	92.9	91.1	87.9	89.47	5.0	1.8	3.43
MUG-U-7B, Qwen2.5-VL-72B, Step1V	92.9	91.1	87.9	89.20	5.0	1.8	3.70
Ovis2-4B, Qwen2.5-VL-72B, Step1V	92.9	90.9	87.9	89.13	5.0	2.0	3.77
Ovis2-4B, Qwen2-VL-72B-Instruct, Step1V	92.9	90.9	88.6	89.43	4.3	2.0	3.47
Ovis2-4B, Qwen2-VL-72B-Instruct, SenseChat-Vision	92.9	90.9	88.8	89.70	4.1	2.0	3.20
MUG-U-7B, Ovis2-1B, Step1o	92.9	92.6	89.0	90.90	3.9	0.3	2.00
InternVL2-76B, Qwen2-VL-72B-Instruct, Step1o	92.8	92.6	84.2	88.53	8.6	0.2	4.27
Qwen2-VL-7B-Instruct, SenseChat-Vision, Step1o	92.8	92.6	84.3	88.77	8.5	0.2	4.03
MUG-U-7B, Qwen2-VL-7B-Instruct, Step1V	92.8	91.1	84.3	88.00	8.5	1.7	4.80
Qwen2-VL-7B-Instruct, Step1o, bailingMM-Lite	92.8	92.6	84.3	87.37	8.5	0.2	5.43
NVLM, SenseChat-Vision, Step1o	92.8	92.6	84.9	88.97	7.9	0.2	3.83
NVLM, Ovis2-8B, Step1o	92.8	92.6	84.9	88.93	7.9	0.2	3.87
MUG-U-7B, MiniCPM-V-2_6, Step1o	92.8	92.6	85.2	89.63	7.6	0.2	3.17
MUG-U-7B, Step1V, bailingMM-Lite	92.8	91.1	85.2	88.30	7.6	1.7	4.50
Ovis2-4B, Qwen2.5-VL-72B, bailingMM-Lite	92.8	90.9	85.2	88.00	7.6	1.9	4.80
InternVL2_5-78B, Step1V, Step1o	92.8	92.6	85.3	88.83	7.5	0.2	3.97
Ovis2-4B, Qwen2.5-VL-7B, Step1V	92.8	90.9	87.4	88.97	5.4	1.9	3.83
Ovis2-1B, Qwen2.5-VL-7B, Step1V	92.8	89.0	87.4	88.33	5.4	3.8	4.47
Ovis2-8B, Qwen2.5-VL-72B, Step1o	92.8	92.6	87.9	89.93	4.9	0.2	2.87
Ovis2-1B, Qwen2.5-VL-72B, Step1o	92.8	92.6	87.9	89.83	4.9	0.2	2.97
Ovis2-4B, Qwen2.5-VL-72B, SenseChat-Vision	92.8	90.9	87.9	89.40	4.9	1.9	3.40
MUG-U-7B, Ovis2-1B, Ovis2-4B	92.8	91.1	89.0	90.33	3.8	1.7	2.47
Ovis2-1B, Qwen-VL-Plus-0809, Step1o	92.7	92.6	84.3	88.63	8.4	0.1	4.07
Ovis2-1B, Qwen2-VL-7B-Instruct, Step1o	92.7	92.6	84.3	88.63	8.4	0.1	4.07
InternVL2_5-78B, Qwen2-VL-7B-Instruct, Step1o	92.7	92.6	84.3	87.40	8.4	0.1	5.30
NVLM, Qwen2-VL-72B-Instruct, Step1o	92.7	92.6	84.9	88.77	7.8	0.1	3.93
Ovis2-34B, Step1o, bailingMM-Lite	92.7	92.6	85.2	89.17	7.5	0.1	3.53
MUG-U-7B, Qwen2-VL-72B-Instruct, bailingMM-Lite	92.7	91.1	85.2	88.37	7.5	1.6	4.33
Ovis2-4B, Qwen2-VL-72B-Instruct, bailingMM-Lite	92.7	90.9	85.2	88.30	7.5	1.8	4.40

241 J Additional Representation Space Analysis

242 Additional analysis of VLM output convergence and divergence patterns on 210 models (beyond Figure ??) will be released with the code
243 and dataset at <https://github.com/Aslan-yulong/consensus-entropy>.

244 K Limitations and Future Directions

245 Despite the promising results demonstrated by the Consensus Entropy framework, several limitations warrant acknowledgment and suggest
246 avenues for future research. The current implementation relies on semantic embedding spaces that may not fully capture the nuances of
247 specialized domains such as mathematical formulas or chemical notations. Additionally, the framework’s performance is contingent on
248 having multiple independent VLMs available at inference time, which may impose computational constraints in resource-limited scenarios.
249 The optimal threshold for routing decisions currently requires empirical calibration for each specific application domain, limiting immediate
250 out-of-the-box deployment.

251 Future research directions could address these limitations through domain-specific embedding techniques for specialized content types,
252 more efficient ensemble methods requiring fewer models, and adaptive threshold mechanisms that automatically adjust to document
253 characteristics. Additional investigations into the theoretical properties of prediction convergence patterns could lead to more principled
254 frameworks for uncertainty quantification beyond the OCR domain. The insights from these convergence patterns might also inform model
255 design and training objectives, potentially enhancing individual model robustness. Exploring the relationship between model architecture
256 diversity and ensemble effectiveness represents another promising direction, particularly in identifying minimal but complementary model
257 combinations that maximize performance gains while minimizing computational overhead.

258 References

- 259 [1] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings
260 Through Self-Knowledge Distillation. *arXiv:2402.03216* [cs.CL]
- 261 [2] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source
262 toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11198–11201.
- 263 [3] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024. OCRBench: on the hidden
264 mystery of OCR in large multimodal models. *Science China Information Sciences* 67, 12 (Dec. 2024). doi:10.1007/s11432-024-4235-6
- 265 [4] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought
266 reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).
- 267 [5] Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, LianWen Jin, and Junyang Lin. 2024.
268 CC-OCR: A Comprehensive and Challenging OCR Benchmark for Evaluating Large Multimodal Models in Literacy. *arXiv:2412.02210* [cs.CV] <https://arxiv.org/abs/2412.02210>