CVPR JUNE 3-7, 2026 DENVER COLORADO

Consensus Entropy: Harnessing Multi-VLM Agreement for Self-Verifying and Self-Improving OCR

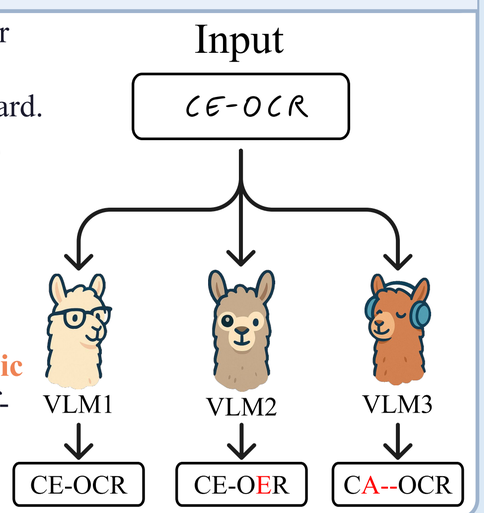
Yulong Zhang [*123], Tianyi Liang [*14], Xinyue Huang [5], Erfei Cui [13], Guoqing Wang [4], Xu Guo [2], Chenhui Li [4], Gongshen Liu [3]

上海创智学院 Shanghai Innovation Institute

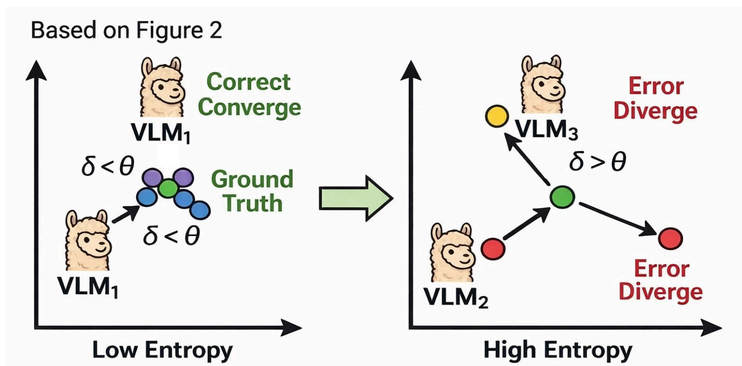
1. Introduction & Motivation

OCR quality is crucial for VLMs, but evaluating it without ground truth is hard. Existing metrics (**BLEU, ROUGE**) fail for OCR.

We need a **agnostic training-free** way to self-verify OCR.

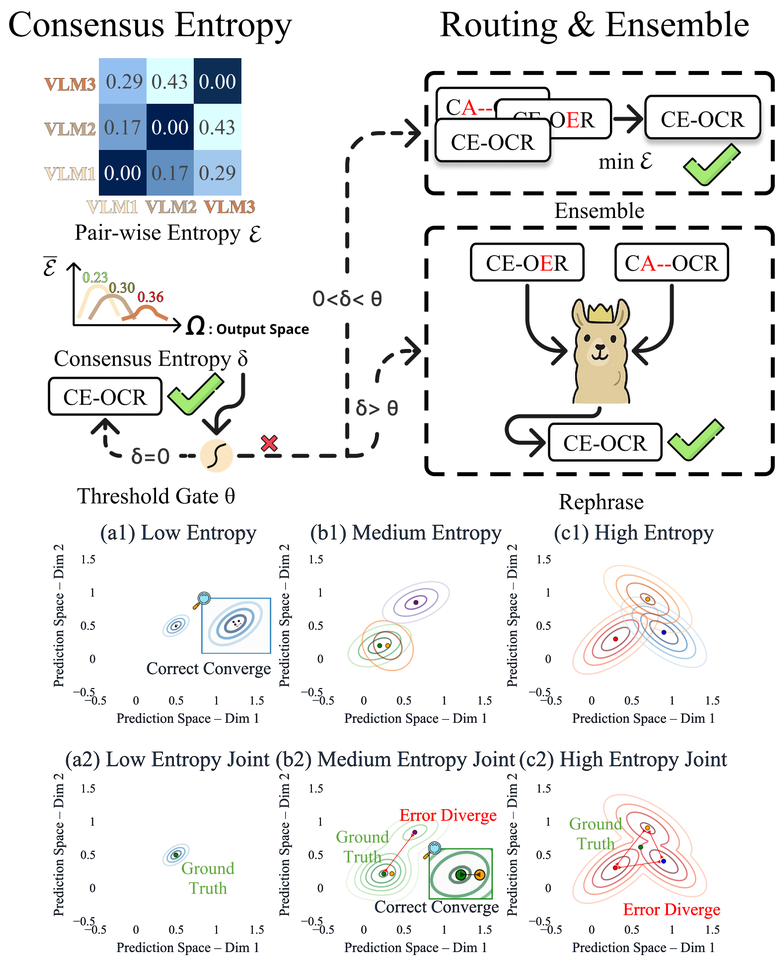


2. Core Concept: Consensus Entropy (CE)



Key Insight: Correct predictions from multiple models **converge** in output space, while erroneous ones **diverge**. CE measures this agreement.

3. CE-OCR Framework



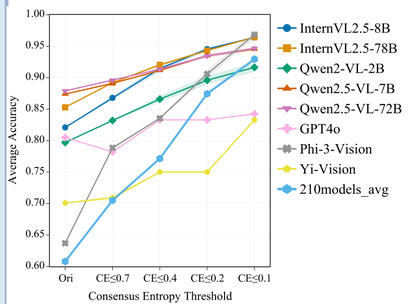
A lightweight, multi-model framework that verifies outputs and **adaptively ensembles** or **routes** to a stronger model for rephrasing based on CE.

4. Results & Performance

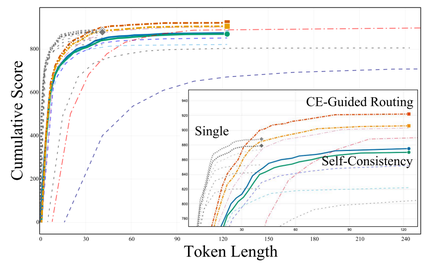
Better than VLM-as-Judge

Human Score Band	GPT4o		Qwen2-VL-7B		Qwen2-VL-72B	
	VLM-J	CE	VLM-J	CE	VLM-J	CE
0.9-1.0	0.7212	0.5729	0.7005	0.7088	0.7227	0.6613
0.7-0.8	0.1842	0.4019	0.1982	0.3956	0.2301	0.4014
0.4-0.6	0.2147	0.3867	0.0000	0.2734	0.1143	0.2847
0.0-0.3	0.4821	0.6544	0.5474	0.6742	0.5283	0.6957
Overall	**0.40**	**0.48 (↑0.08)**	**0.36**	**0.51 (↑0.15)**	**0.40**	**0.51 (↑0.11)**

Self-Verification



Efficiency



Self-Improving

Method	En	Math	Elem	Cn All
GPT4o	61.2	43.4	29.8	32.2
InternVL2.5-26B	65.6	37.4	32.6	44.2
Gemini Pro	61.2	47.7	30.9	43.1
CE-Ensemble	67.2	50.1	**34.0**	45.7
CE-OCR (GPT4o Rephrase)	**71.6**	**53.1**	33.8	**48.0**
Relative Δ vs Ensemble (%)	+6.5%	+6.0%	-0.5%	+5.0%
Relative Δ vs Best Single (%)	+9.1%	+11.3%	+3.7%	+8.6%

Cheap open-source models (< 10B) outperform larger SOTA models

Model Combinations of CE-Ensemble (Individual Scores)		
*Ovis2-1B, Qwen2.5VL-7B, Step1V, **Step1o** (890, 874, 886, **926**)	**CE-Ensemble: 955**	**Gain: +29**
†Ovis2-1B, **Ovis2-4B**, Qwen2VL-7B, Qwen2.5VL-7B (890, **909**, 843, 874)	**CE-Ensemble: 933**	**Gain: +24**
*Ovis2-4B, Qwen2.5VL-7B, **Step1o** (909, 874, **926**)	**CE-Ensemble: 938**	**Gain: +12**
InternVL2.5-78B, Qwen2.5VL-72B, **Qwen2VL-72B** (853, 879, **888**)	**CE-Ensemble: 920**	**Gain: +32**
†InternVL2.5-8B, Qwen2VL-7B, **Qwen2.5VL-7B** (821, 843, **874**)	**CE-Ensemble: 897**	**Gain: +23**

Beyond OCR: VQA Tasks

Category	Baseline	CE-Ensemble (Δ)	Better Metric
Scene-VQA (easy)	92.5	**98.0 (+6.0%)**	Cosine Distance
Doc-VQA (easy)	87.5	**90.5 (+3.4%)**	Edit Distance
Formula (easy)	82.0	**88.0 (+7.3%)**	Edit Distance
Science-VQA	61.3	**63.7 (+3.9%)**	Cosine Distance
Math-VQA	40.0	**45.6 (+14.0%)**	Cosine Distance
Knowl.-Reason.	60.3	**66.3 (+10.0%)**	Cosine Distance
Visual-Und.	75.7	**82.4 (+8.9%)**	Cosine Distance

Arxiv

Github

https://github.com/Aslan-yulong/consensus-entropy

Training-free. Model-agnostic. High performance. Low cost.