Visual-ERM: Reward Modeling for Visual Equivalence

Ziyu Liu*, Shengyuan Ding*, Xinyu Fang,
Xuanlang Dai, Penghui Yang, Jianze Liang,
Jiaqi Wang, Kai Chen, Dahua Lin, Yuhang Zang†

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

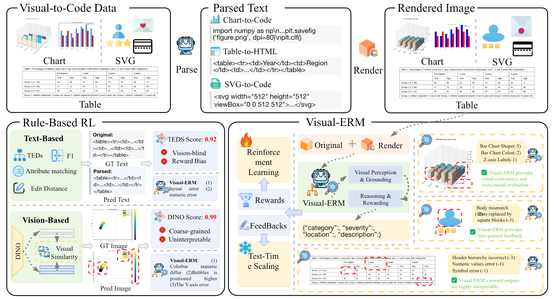
Shanghai Jiao Tong University · Shanghai AI Laboratory · Fudan University · CUHK

Why visual rewards?

The task. Vision-to-code reconstructs charts, tables and SVGs as executable code or markup.

Reward mismatch. Text metrics miss alignment, spacing, geometry and color. Embedding similarity can reward semantically similar outputs despite local errors.

Visual-ERM. Compare the reference and rendered prediction, grounding judgment in both visual structure and embedded text.



Structured feedback. Each report specifies category, severity, location and description, supporting interpretable judgment across all three tasks.

Chart-to-Code: stronger policies

Model	ChartMimic [41] (Direct)				ChartMimic [41] (Customized)				Avg↑
	Exec_rate↑	Low↑	High↑	Overall↑	Exec_rate↑	Low↑	High↑	Overall↑	
Baseline									
InternVL3.5-8B [38]	62.2	41.9	48.9	45.4	79.7	53.0	63.3	58.2	51.8
Qwen3-VL-8B-Instruct [3]	82.2	62.7	72.7	67.7	86.8	66.3	76.8	71.6	69.6
VinciCoder-8B-SFT [47]	86.2	68.2	77.5	72.9	86.2	52.3	71.5	61.9	67.4
JanusCoderV-8B [36]	80.6	65.8	73.2	69.5	80.7	66.7	74.2	70.4	70.0
ChartMaster-7B [37]	93.3	74.5	82.1	78.3	88.8	59.5	74.2	66.8	72.6
Qwen3-VL-8B-Instruct [3]									
+ RL (DINO-based Reward)	90.8	71.1	81.9	76.5	91.7	71.2	80.3	75.8	76.1
+ RL (Visual-ERM)	92.5	74.4	84.6	79.5	91.5	71.8	81.1	76.5	78.0
Δvs. Qwen3-VL-8B-Instruct	+10.3	+11.9	+11.9	+11.8	+4.7	+5.5	+4.3	+4.9	+8.4
VinciCoder-8B-SFT [47]									
+ RL (DINO-based Reward)	90.8	73.1	82.3	77.7	91.5	57.0	77.5	67.3	72.5
+ RL (Visual-ERM)	94.3	78.0	88.5	83.2	95.0	60.9	82.6	71.7	77.5
Δvs. VinciCoder-8B-SFT	+8.1	+9.8	+11.0	+10.3	+8.8	+8.6	+11.1	+9.8	+10.1

Consistent gains. Average scores improve by +8.4 for Qwen3-VL-8B and +10.1 for VinciCoder. Both outperform their DINO-based RL variants.

NeurIPS 2026

Visual-ERM

Reward Modeling for Visual Equivalence

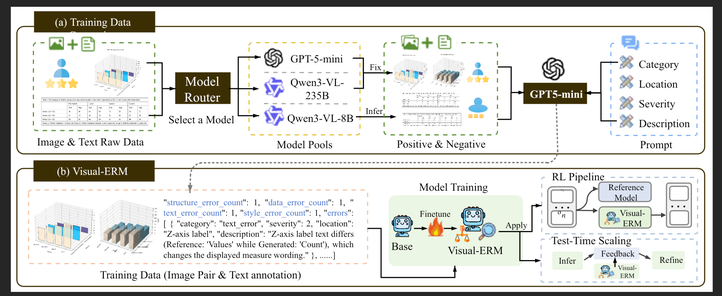
Visual equivalence. An 8B generative critic jointly inspects global structure, local geometry and rendered text.

RL supervision. Predicted error severities become scalar rewards that guide the policy toward faithful renderings.

Inference feedback. Localized discrepancy descriptions support reflection and revision, making each judgment interpretable.

Data generation and critic training

1. Construct pairs. Combine targeted code edits with weaker-model predictions to capture controlled errors and natural failures.
2. Annotate discrepancies. GPT-5-mini compares each image pair and produces a structured error report.
3. Train the critic. Fine-tune Qwen3-VL-8B-Instruct to generate reports with a negative log-likelihood objective.



From discrepancy reports to rewards

Severity aggregation. Sum error severities and normalize within each task batch. Subtract the normalized score from one to obtain a bounded fidelity reward.

Render success. Add a success term to encourage executable outputs alongside visually faithful reconstructions.

Policy optimization. GRPO uses the combined reward with a reference-model KL constraint. The same design supports multiple vision-to-code tasks.

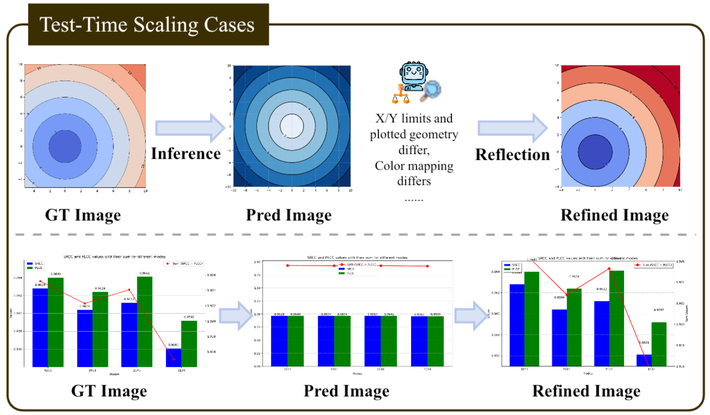
VC-RewardBench

Dataset. 1,335 reference/corrupted image pairs cover charts, tables and SVGs. Human reviewers consolidate preliminary annotations from multiple models.

Evaluation. Match predicted discrepancies against reference annotations to measure precision, recall and F1. Severity correlation captures agreement on overall error scores.

Findings. Visual-ERM reaches hard / soft F1 of 42.1 / 44.7, versus 29.5 / 32.4 for Qwen3-VL-235B. Specialized supervision strengthens judgment at a smaller model scale.

Test-Time Scaling



Refinement loop. Generate and render code, then revise it using the previous prediction and localized feedback. The examples show corrections to geometry and color mapping.

Three-round results. ChartMimic improves from 69.6 to 77.6; the RL-tuned model improves from 78.0 to 81.1.

Table-to-Markdown: faithful structure

Model	OmniDocBench [26]			OmniDocBench-v1.5 [26]			olmOCRBench [28]	Avg↑
	TEDS↑	TEDS-S↑	Edit-Dist↓	TEDS↑	TEDS-S↑	Edit-Dist↓	TA↑	
Baseline								
Qwen2.5-VL-7B [4]	74.4	80.1	58.9	69.3	74.1	58.7	71.1	64.5
InternVL3.5-8B [38]	71.0	77.6	44.9	62.9	70.4	44.1	74.1	66.7
Qwen3-VL-8B-Instruct [3]	78.9	83.9	23.2	72.7	77.2	26.9	75.3	76.8
Qwen3-VL-8B-Instruct [3]								
+ RL (DINO-based Reward)	62.2	69.5	37.0	61.1	67.4	37.9	71.7	65.3
+ RL (TEDS-based Reward)	79.2	82.9	31.6	73.0	76.6	35.3	**78.6**	74.8
+ RL (Visual-ERM)	**81.4**	**86.3**	**20.7**	**75.4**	**80.4**	**24.2**	78.1	**79.5**
Δvs. Qwen3-VL-8B-Instruct	+2.5	+2.4	+2.5	+2.7	+3.2	+2.7	+2.8	+2.7

Average: 76.8 → 79.5. Both text and structure metrics improve.