



Stop Thinking, Start Looking

Efficient Post-Training for Multimodal Document QA via Reasoning-Free Alignment

Harikrishnan P M · Goutham Vignesh · Ganesh Parab · Saisubramaniam Gopalakrishnan · Vishal Vaddina · Varun V · Rohit Agrawal

Phi Labs, Quantiphi

62%

per-query inference
token reduction

65%

less training data
(early SFT→RL)

0.718

ID F1_all with
Perception-RFT

Problem Statement

Document Visual Grounding (DVG) requires the model to jointly extract the correct answer text AND return its precise bounding box on the document image.

MEGA ELECTRONICS CENTRE
LOT 12-B, JALAN DAMAI 3,
TAMAN DAMAI JAYA
47650 BANDAR BARU, SELANGOR
Tel: +603-8912 5677
GST ID No: 001273896541

-- TAX INVOICE --

INVOICE NO: EL-5781/09/T0421
Date: 25-04-2022 Time: 10:23 AM
Cashier: PARAH

1x SAMSUNG USB CABLE TS-330	RM 25.90	SR
2x POWER ADAPTER 2.1A	RM 39.80	SR
@RM 19.90		

Total Sales (Inclusive GST)	RM 65.70
CASH	RM 70.00
CHANGE	RM 4.30

GST Summary	Amount (RM)	Tax (RM)
SR @ 6%	61.98	3.72
Total	61.98	3.72

Thank You & Please Come Again!
Goods sold are not returnable or exchangeable
TERIMA KASIH SILA DATANG LAGI
Please keep receipt for warranty claims

Question: Extract the company address
Predicted Answer: LOT 12-B, JALAN DAMAI 3, TAMAN DAMAI JAYA
Predicted Location: [56, 39, 260, 79]

SFT Saturates

Scales with annotated data but hits optimization plateaus, more data yields diminishing geometric precision gains.

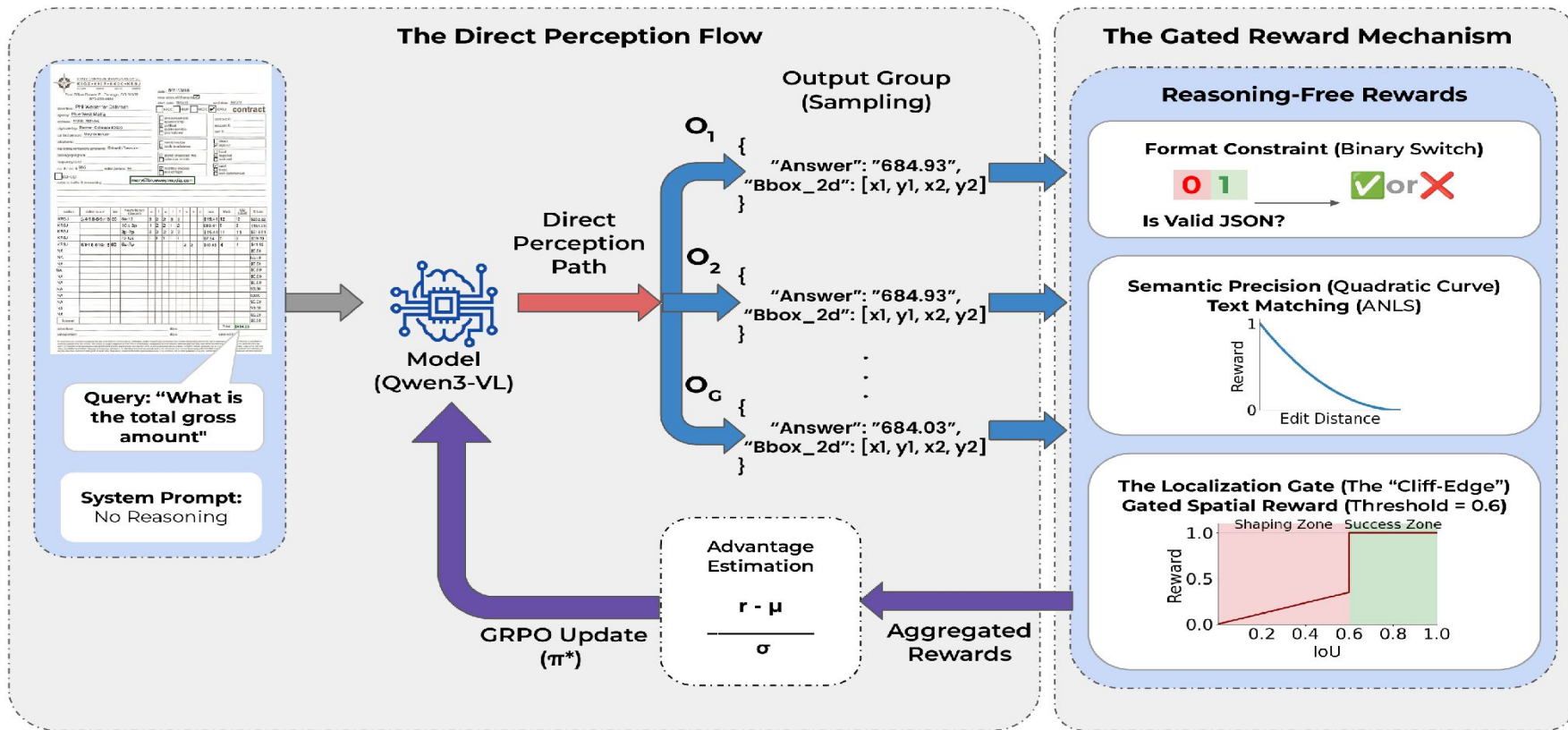
VS

Reasoning RL is Expensive

Verbose <think> traces inflate inference token cost by 60%+ with no clear benefit for inherently perceptual grounding tasks.

Hypothesis: DVG is a perception-dominant task, the optimal policy looks and localizes directly, without narrating reasoning.

Proposed Approach



Perception-RFT: Apply GRPO under a Direct Perception constraint

Experimental Setup

Model & Infrastructure

- Qwen3-VL-4B foundation model
- LoRA adaptation ($r=16$, $\alpha=16$)
- GRPO for reinforcement fine-tuning
- Single NVIDIA A100 80GB + Unsloth

Dataset Construction

- 23,696 QA samples (DocILE + FormNLU)
- Financial docs: invoices, POs, receipts
- Semantic augmentation + phrasing variation
- Dynamic sub-sampling, no layout memorization

Evaluation Benchmarks

- ID: 6,194 finance samples (held-out)
- OOD 1: DOGR-Bench, 800 samples
- OOD 2: MMDocBench, 4028 samples
- Metrics: F1_EM, F1_loc, F1_all

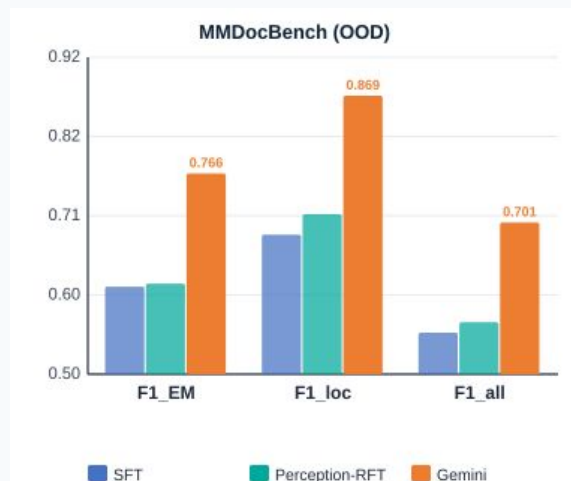
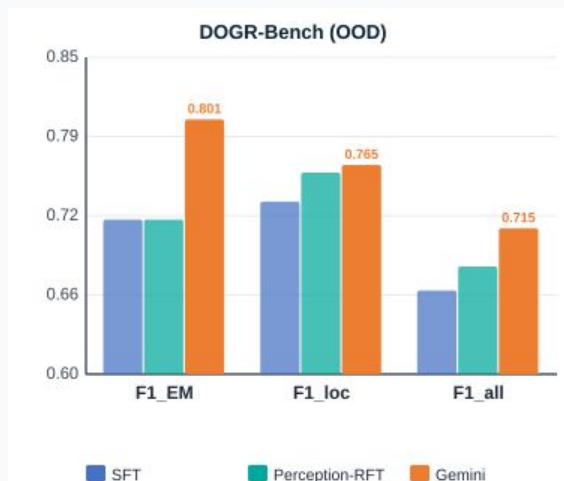
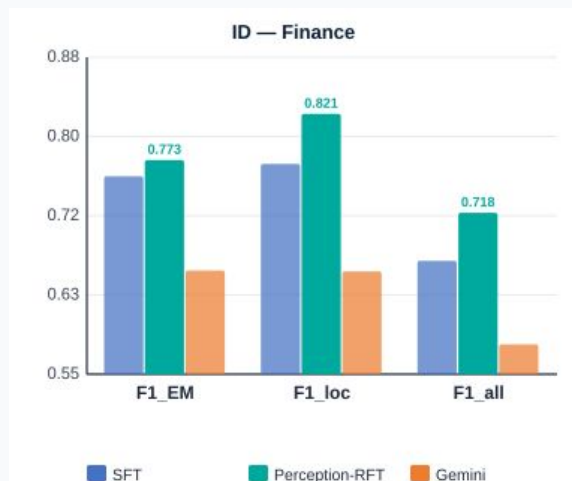
Compared Setups

- Zero-shot
- SFT: 3 epochs
- RFTs: SFT → RL (6k steps)
- RFTb: cold-start RL
- Reasoning-RFTb
- Gemini 3.0 Flash

Controlled comparison of alignment paradigms

Key control: Reasoning-enabled RL uses the identical reward function as Perception-RFT, differences isolate the value of intermediate reasoning traces, not reward design.

Main Results



RFTs leads across all ID metrics

- F1_EM (0.773), F1_loc (0.821), F1_all (0.718), All three improve over SFT (0.756 / 0.769 / 0.668).
- RL refines both semantic extraction and geometric localization.



Grounding Divergence: F1_loc transfers, F1_EM lags

- DOGR:
F1_loc 0.416 → 0.759 (+34pt)
F1_EM 0.743 → 0.722 (-2pt)
- Geometry generalizes across domains; semantics over-specializes to training distribution.

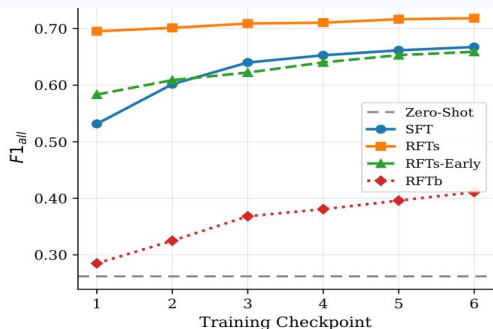


Specialist 4B beats Gemini in-domain; Gemini leads broad OOD

- ID: RFTs F1_all 0.718 vs Gemini 0.581 (+24%).
- MMDoc: Gemini 0.701 vs RFTs 0.569.
- Task alignment > scale in-domain.
- Generalist breadth matters beyond training distribution.

Takeaway & Conclusion

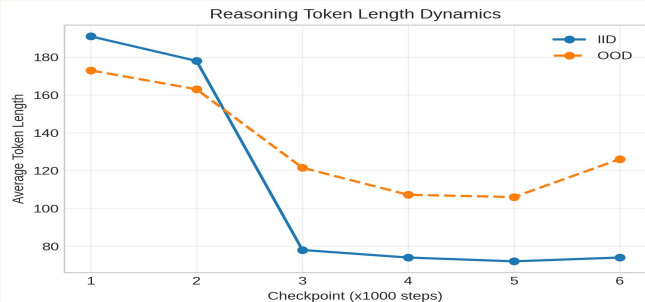
1 RL breaks SFT saturation



On in-domain data, SFT plateaus while RFTs continues to improve semantic and localization quality

An early SFT→RL transition reaches comparable precision with ~65% less training data

2 Reasoning hurts and then disappears



Under identical RL settings, RFTb outperforms Reasoning-RFTb:

- ID F1all 0.411 vs 0.303
- OOD F1all 0.600 vs 0.382

During training, reasoning traces compress from ~191 to ~72 tokens on ID, a 62% reduction

3 Grounding Divergence

Under OOD shift, localization transfers strongly, but semantic extraction can regress



This reveals an asymmetric trade-off: geometry improves more reliably than semantic robustness

Core contribution: Perception-RFT directly optimizes structured grounding outputs without paying for intermediate reasoning traces, achieving superior precision, cheaper inference, and better data efficiency.

"For DVG, stop thinking tokens; start looking at the document."