

Not All Tokens Should Be Treated Equally: Context Credits Reassignment

Shujian Gao^{1,2,*} Yuan Wang^{4,*} Jiamei Yan³ Penghao Zhou^{3,†} Qinglei Wang^{3,†} Jiangtao Yan⁵ Zuxuan Wu^{1,2,†} Yu-Gang Jiang^{1,2,†}

¹Institute of Trustworthy Embodied AI, Fudan University ²Shanghai Innovation Institute ³ByteDance TikTok ⁴Zhejiang University ⁵Wuhan University

*Equal contribution †Project lead ‡Corresponding authors



NeurIPS 2026

Poster · Code & paper via QR →

QR
paper

QR
code

TL;DR

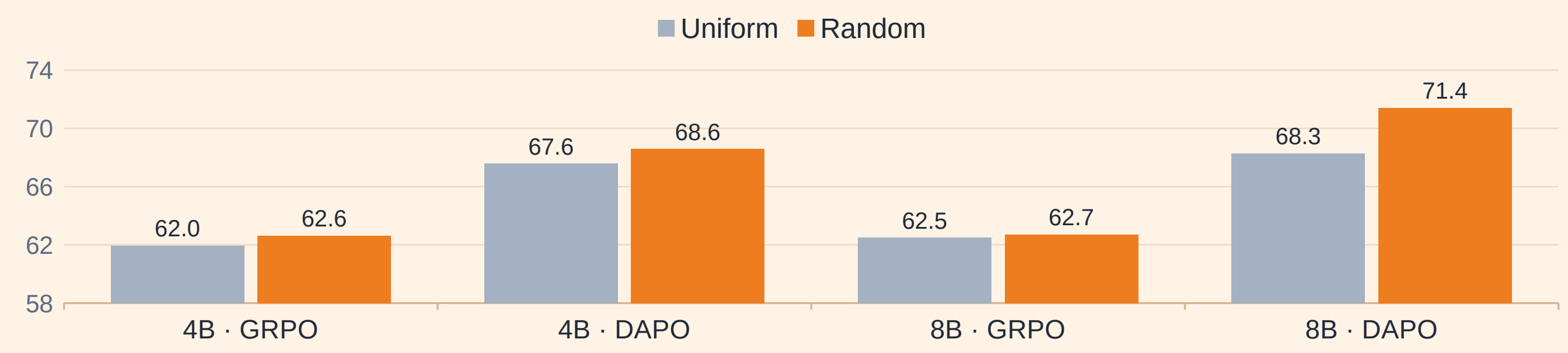
Multimodal RLVR gives **every token the same advantage**, which quietly **weakens visual grounding**. **HiCCR** reassigns outcome credit to **visually-grounded tokens & trajectories** using intra-visual attention concentration — **up to +8.0 AVG@8** over GRPO, **<3% overhead**, no extra model or data.

1 Is Uniform Credit Right?

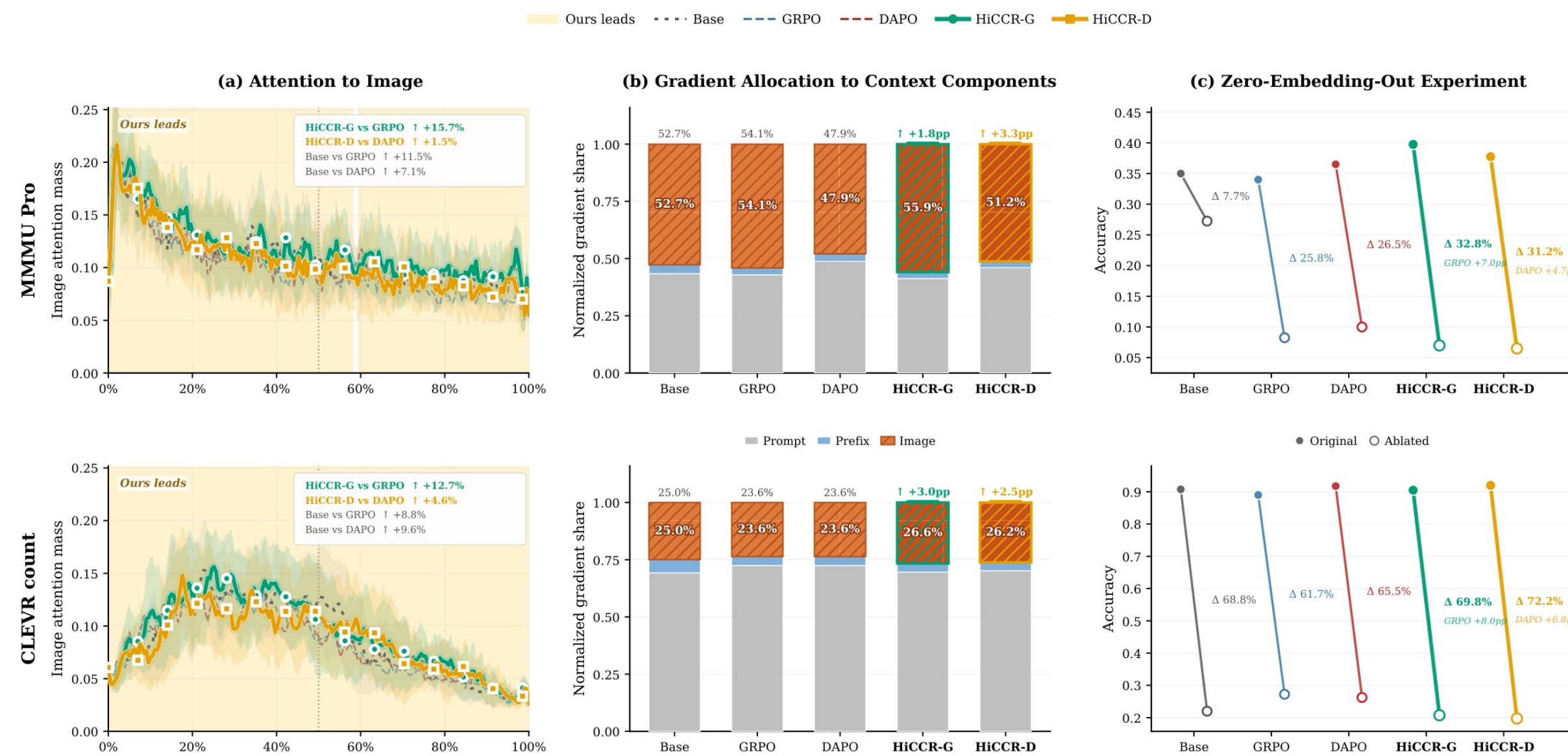
GRPO / DAPO broadcast **one scalar advantage** to every token, yet tokens differ greatly in how much they **rely on visual evidence**.

$$\nabla_{\theta} \mathcal{L}_t \propto \hat{A} \nabla_{\theta} \log \pi_{\theta}(y_t | x, y_{<t})$$

Sanity check: scaling advantages by **random noise** $w_t \sim N(1, 0.2^2)$ already **beats uniform credit** in 4 / 4 settings.



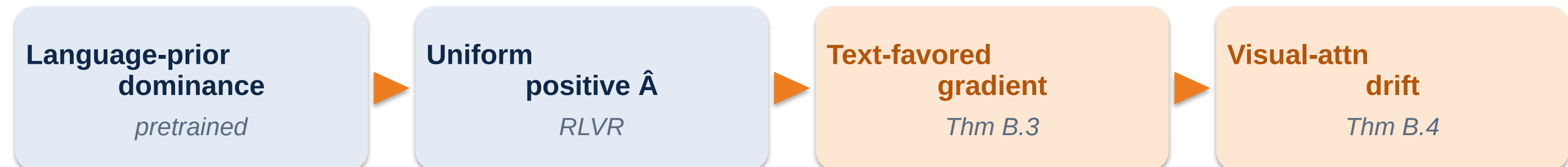
2 Finding: RLVR Weakens Vision



- ▶ (a) **Attention:** RLVR models attend less to the image; gap widens over generation
- ▶ (b) **Gradient:** smaller Input×Gradient share routed to image tokens
- ▶ (c) **Function:** zeroing image embeddings hurts RLVR models less

✓ HiCCR mitigates all three

3 Why? A Theoretical View

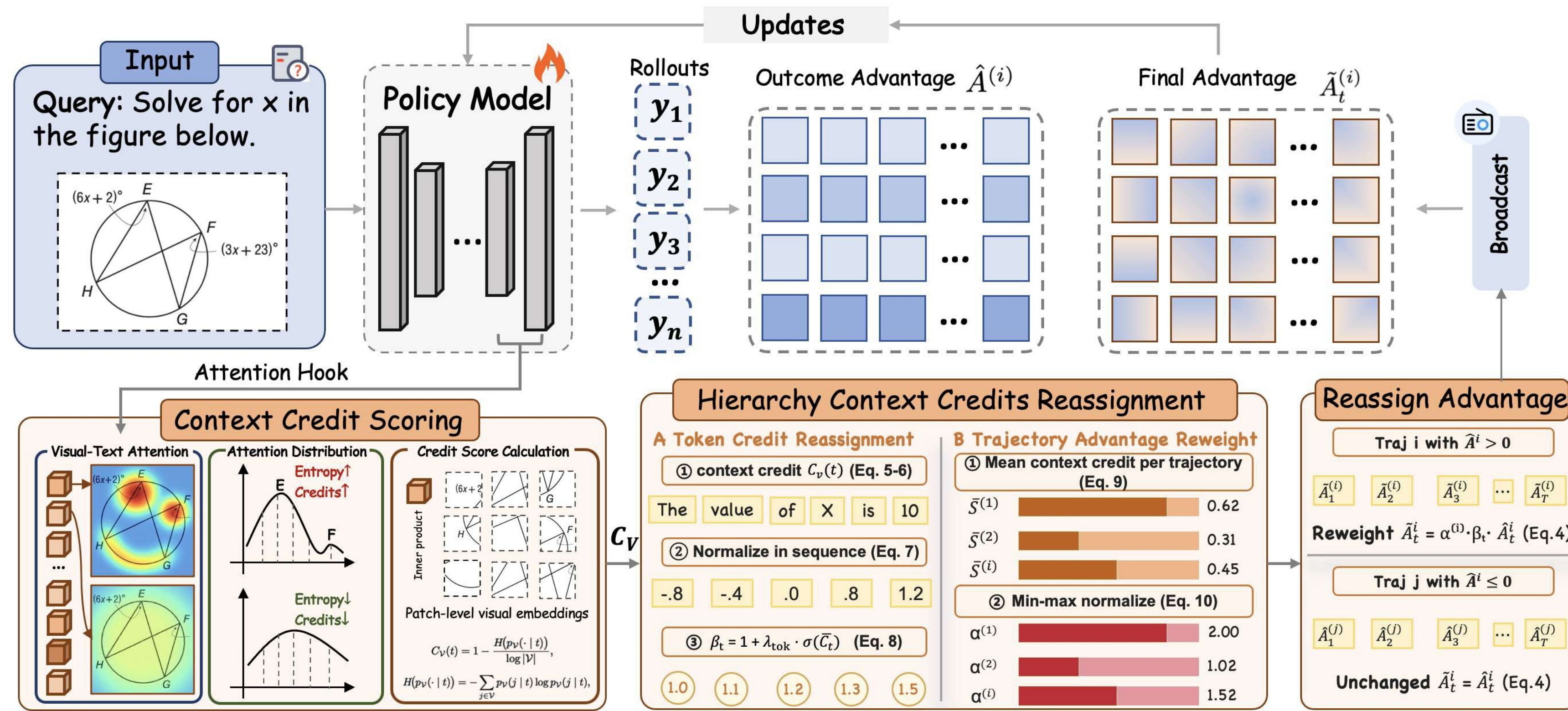


Thm B.3 Uniform credit lets the textual gradient dominate the visual one:

$$\frac{\|G_t^x\|_F}{\|G_t^y\|_F} \geq \frac{c_T c_J}{C_J} \rho_t > 1$$

Thm B.4 The bias accumulates into attention drift away from image tokens: RLVR **amplifies** rather than creates it.

4 HiCCR: Hierarchical Context Credit Reassignment



Reassigned advantage only when $\hat{A} > 0$; negative rollouts unchanged

$$\tilde{A}_t^{(i)} = \alpha^{(i)} \cdot \beta_t^{(i)} \cdot \hat{A}^{(i)}$$

① Context Credit Score

Visual-attention concentration: 1 – normalized entropy of attention restricted to image tokens (last 2 layers, head-avg). Forward-hook only.

$$p_{\mathcal{V}}(j|t) = \frac{\exp(e_{t,j}/\tau)}{\sum_{j' \in \mathcal{V}} \exp(e_{t,j'}/\tau)} \quad C_{\mathcal{V}}(t) = 1 - \frac{H(p_{\mathcal{V}}(\cdot|t))}{\log|\mathcal{V}|} \in [0, 1]$$

② Token Credit Reassignment

Amplifies tokens with above-average visual concentration. No token suppressed, no sign flip.

$$\bar{c}_t = \frac{C_{\mathcal{V}}(t) - \mu_c}{\sigma_c + \varepsilon}, \quad \beta_t = 1 + \lambda_{\text{tok}} \sigma(\bar{c}_t) \in (1, 1 + \lambda_{\text{tok}})$$

③ Trajectory Advantage Reweight

Among correct rollouts, upweights those with stronger visual engagement; $\alpha \geq 1$ keeps RL stable.

$$\bar{S}^{(i)} = \frac{1}{T_i} \sum_t C_{\mathcal{V}}^{(i)}(t), \quad \alpha^{(i)} = 1 + \frac{\bar{S}^{(i)} - \min_k \bar{S}^{(k)}}{\max_k \bar{S}^{(k)} - \min_k \bar{S}^{(k)} + \varepsilon}$$

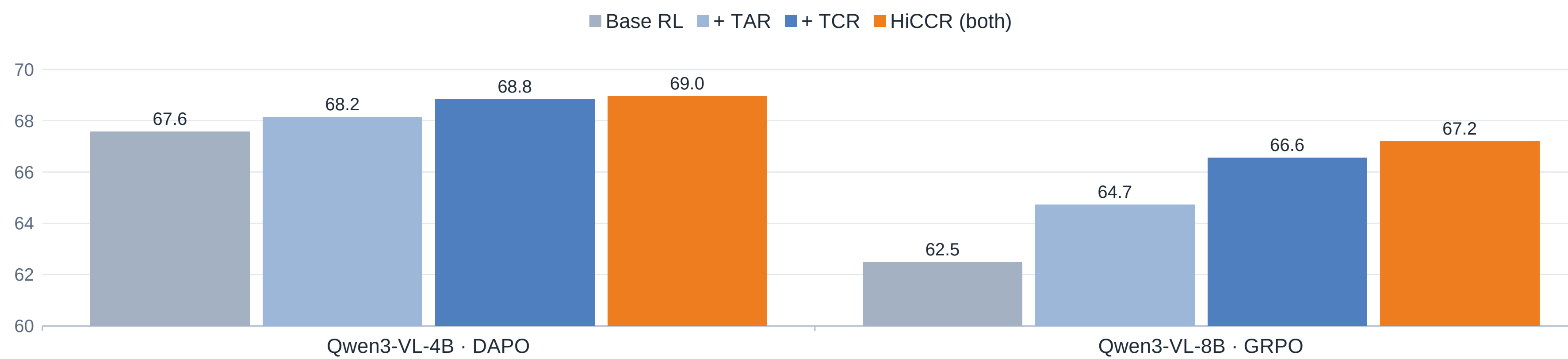
Drop-in
for GRPO / DAPO

+2.7% time
+4.5% memory

No extra model
or data

No extra
backward pass

Ablation: token & trajectory levels are complementary



5 Results on 10 Benchmarks

	Qwen3-VL-4B			Qwen3-VL-8B		
Method	Math	General	AVG@8	Math	General	AVG@8
Base (Instruct)	49.78	48.17	49.07	56.33	54.94	55.71
Best other RLVR†	69.60	61.95	66.20	74.94	69.95	72.32
GRPO	64.17	59.17	61.95	64.24	60.31	62.49
HiCCR-G (ours)	72.65	66.55	69.94 ^{+7.99}	69.79	63.94	67.19 ^{+4.70}
DAPO	71.40	62.80	67.58	70.95	64.96	68.29
HiCCR-D (ours)	71.36	65.93	68.95 ^{+1.37}	76.01	68.69	72.76 ^{+4.47}

† strongest of 7 other RLVR methods per backbone (DVPR for 4B, Reinforce++ for 8B); superscripts = gain over matched base RL

72.76

8B HiCCR-D AVG@8
> Qwen2.5-VL-72B (61.65)

+7.99

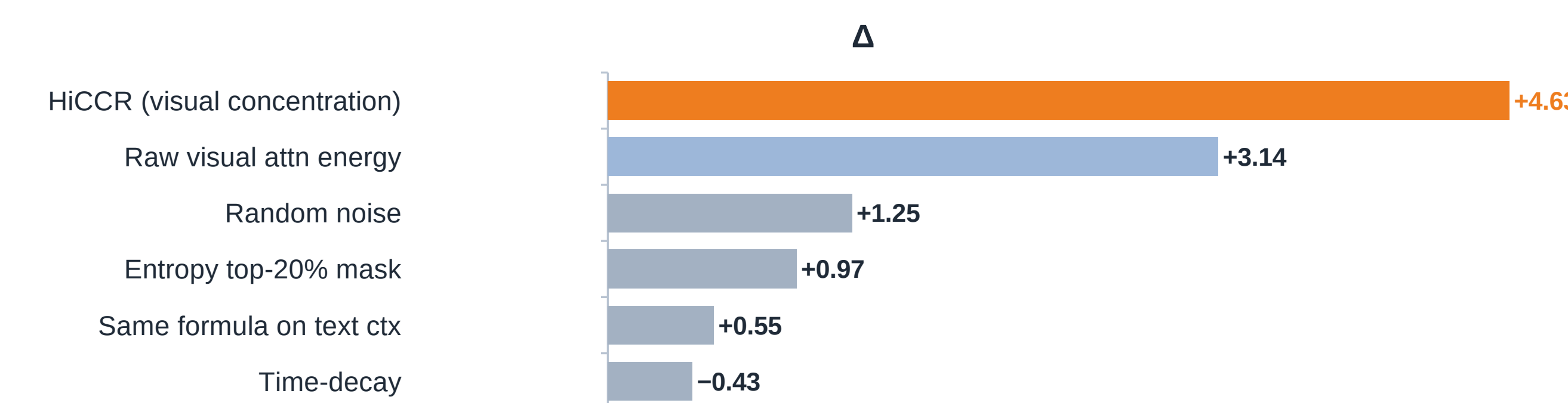
over GRPO
Qwen3-VL-4B

72.2

MMStar (8B)
best - also HalluBench

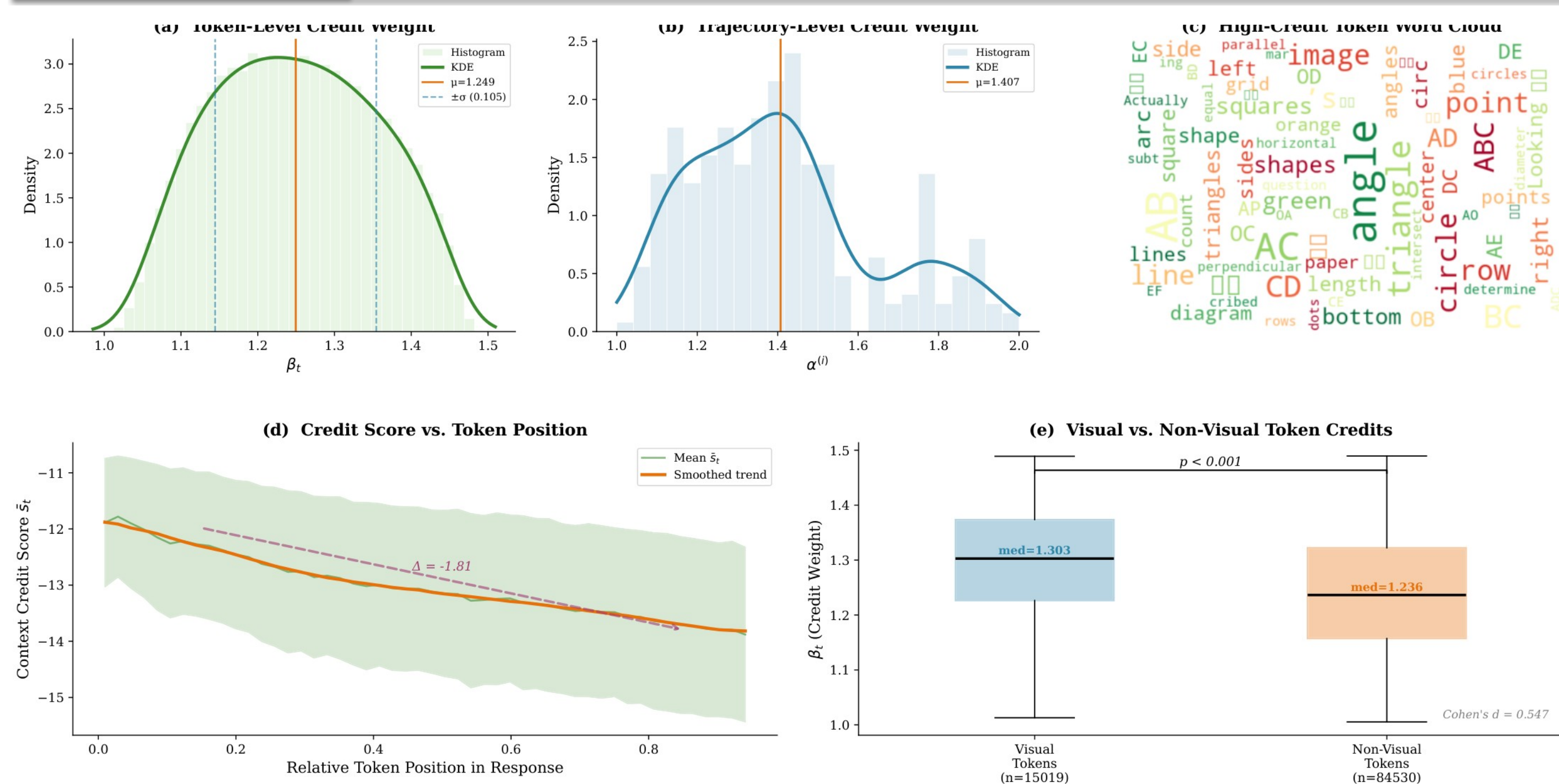
What makes a good credit signal?

Δ AVG@8 vs. uniform credit, averaged over {4B, 8B} × {GRPO, DAPO}



→ The operative signal is **concentration within the visual subset** — not mere non-uniformity, text context, or attention magnitude.

6 Where Does the Credit Go?



Visual tokens get higher β ($p < 0.001$); top-credit words are **angle**, **circle**, **triangle**, **diagram**; credit is front-loaded while the reasoning scaffold is built.