Bridging the Perceptual Gap: Rule-Grounded Representations for Tactical Reasoning in Multi-Sport Video Understanding

Sree Krishna Vadde

Abstract

Multimodal large language models (MLLMs) for sports understanding can describe actions but remain fundamentally unreliable for adjudicating tactical scenarios and rule violations. We identify this as a *Perceptual Gap*: a failure to map perceived visual evidence into the formal predicates required for rule-based reasoning. Diagnostic analysis on the SportR benchmark [9] reveals that 62% of MLLM errors concentrate at this grounding stage. We propose **RuleGround**, a modular neuro-symbolic architecture with an explicit *Rule Grounding Module* (RGM) trained via weak supervision from human-authored rationales and optimized using Group Relative Policy Optimization (GRPO) [3] with Risk-aware Stepwise Alignment (RSA) [13]. In experiments on SportR-Hard, RuleGround improves grounding IoU (Q5) by +4.18 absolute over the Qwen-VL-7B baseline while reducing false-positive infraction predictions by 34% via token-level risk constraints at matched recall.

1 Introduction: The Sports Intelligence Pyramid

Sports understanding requires more than fluent narration: reliable adjudication requires linking visual evidence to formal rule structure, including temporal ordering, spatial relations, and exception conditions. We conceptualize sports intelligence as a three-tier hierarchy:

- **Tier 1: Perceptual Foundation.** Detect entities and local events (ball state, contact, possession).
- **Tier 2: Tactical Grounding.** Map evidence to rule predicates and apply rules to adjudicate infractions.
- **Tier 3: Strategic / Counterfactual Reasoning.** Explain tactical meaning and evaluate alternatives.

Benchmarks such as SportR [9] and SPORTU [10] indicate that Tier 2 remains a major bottleneck even for strong MLLMs: models can often identify salient events but fail to ground them into predicates required for rule-faithful decisions.

Perceptual Gap Hypothesis. We define the *Perceptual Gap* as a systematic failure in mapping perceptual evidence to rule predicates. In a manual review of 200 GPT-4o errors on SportR-style prompts, we attribute 62% to grounding failures: the model mentions relevant evidence (e.g., contact) but misclassifies its rule-relevant status (timing, legality, or exception applicability). This motivates an explicit predicate bottleneck rather than implicit pattern-based grounding.

Contributions.

- **C1 (Architecture):** RuleGround introduces a Rule Grounding Module (RGM), an explicit neuro-symbolic bottleneck that predicts a predicate state and composes rule constraints via differentiable logic.

- **C2 (Supervision):** We propose a weak-supervision pipeline that extracts structured predicate labels from human rationales using a constrained predicate extractor, with reliability estimated via cross-model agreement and auditing (Appendix).
- **C3 (Alignment):** We combine GRPO [3] with RSA [13] to reduce tail-risk failure modes such as false-positive infraction calls via token-level risk constraints.

2 Related Work

2.1 MLLMs and Sports Benchmarks

SportR [9] evaluates multi-sport reasoning with explicit grounding (Q5). SPORTU [10] expands sports understanding evaluation across modalities and reasoning difficulty. These benchmarks collectively show that grounding is a core limitation: models can produce plausible explanations while failing to localize or correctly interpret evidence.

DeepSport [14] employs agentic RL to improve video reasoning pipelines (including dynamic evidence selection). RuleGround differs by enforcing *explicit predicate-level grounding* as a bottleneck and aligning predicate faithfulness to reduce unsupported infractions.

2.2 Modern Temporal Transformers for Video

ActionFormer [12] introduced local self-attention for temporal action localization, achieving strong results on THUMOS and ActivityNet. Recent extensions [7] modernize the architecture with primitives from large language models: Flash Attention [1] for $O(T)$ memory, Rotary Position Embeddings (RoPE) [5, 6] for length generalization, RMSNorm [11] for efficient normalization, and SwiGLU activations [4]. These components directly inform our RGM design. Additionally, SnapFormer [7] adapts ActionFormer for point-event detection via heatmap regression—relevant for detecting instant predicates such as contact moments or ball release timing.

2.3 Risk-Aware RL Alignment

GRPO [3] enables efficient RL post-training via group-relative advantages without requiring a learned critic. RSA [13] extends constrained policy optimization with nested risk measures to suppress rare-but-high-impact errors, which is particularly relevant in adjudication settings where false positives are operationally costly.

3 Error Attribution Protocol

To validate the Perceptual Gap hypothesis, we developed a systematic attribution protocol and applied it to 200 GPT-4o errors on SportR-style adjudication prompts. Two annotators independently assigned each error to one of three categories, with disagreements resolved by discussion. Inter-annotator agreement was substantial (Cohen's $\kappa = 0.71$).

The dominance of grounding errors motivates explicit predicate supervision rather than relying on implicit pattern learning.

4 Problem Setup and Perceptual Gap Formalization

Let v denote a sports clip and y the required judgment (infraction label, foul class, etc.). Let $\mathbf{p} \in [0,1]^K$ be a predicate state over an ontology $\mathcal{P} = \{p_1, \ldots, p_K\}$.

Table 1: Error attribution criteria and observed distribution (200 errors).

Category	Attribution Criteria	% Errors
Perception	Model fails to detect the relevant visual evidence (misses contact event, wrong player, incorrect ball state). Annotators verify that the key entity/action is absent or contradicted by the clip.	24%
Grounding	Model mentions the evidence but misclassifies its *rule-relevant status* (e.g., correct contact but wrong timing relative to release/arrival; correct position but wrong legal state; misses exception predicates).	**62%**
Reasoning	Predicates are correctly grounded, but logical deduction fails (e.g., correct timing predicate but concludes wrong foul type or wrong penalty mapping).	14%

We decompose the pipeline into:

$$v \xrightarrow{P} \{f_t\}_{t=1}^{T} \xrightarrow{G} \hat{\mathbf{p}} \xrightarrow{T} \hat{y},$$

where P is perceptual encoding, G is grounding (predicate estimation), and T is reasoning over predicates for the task output. The Perceptual Gap corresponds to error mass concentrated in failures of G even when P captures relevant evidence (as in Table 1).

5 Method: The RuleGround Architecture

RuleGround inserts a predicate bottleneck between a frozen video encoder and a reasoning head.

5.1 Encoder and Feature Extraction

We use a frozen VideoMAE-v2 ViT-B encoder [8]. Each clip is represented by $T = 16$ uniformly sampled frames, producing frame embeddings $\{f_t\}_{t=1}^{T}$.

5.2 Rule Grounding Module (RGM)

RGM predicts a predicate vector $\hat{\mathbf{p}}$ from video embeddings. For each predicate p_i, we use a 2-layer MLP atop temporal attention pooling:

$$\hat{p}_i = \sigma\left(W_i \cdot \text{pool}(f_1, \ldots, f_T) + b_i\right), \tag{1}$$

where $\text{pool}(\cdot)$ is temporal attention pooling implemented using the v2 transformer block from [7], incorporating RoPE for position encoding and Flash Attention for memory-efficient processing. For instant-event predicates (e.g., `contact_occurred`, `ball_release`), we optionally employ SnapFormer-style heatmap heads [7] to produce frame-level predicate activations before temporal aggregation.

Why a predicate bottleneck? (1) It provides a *structured intermediate representation* aligned with rulebooks. (2) It supports *error attribution* (perception vs. grounding vs. reasoning). (3) It enables *structural transfer*: shared predicates and rule templates can generalize across sports.

5.3 Differentiable Rule Composition

Rules are represented as logical formulas over predicates using differentiable t-norm approximations [2]:

$$\widetilde{\wedge}(a,b) = ab, \quad \widetilde{\vee}(a,b) = a + b - ab, \quad \widetilde{\neg}(a) = 1 - a. \tag{2}$$

Composed rule scores can be used as auxiliary supervision and as structured features for the reasoning head.

5.4 Reasoning Head

The reasoning head consumes the predicate state (and optionally composed rule scores) to produce task outputs (Q1/Q2/Q5). We use a lightweight transformer that maps $\hat{\mathbf{p}}$ plus pooled visual features to $\hat{y}$, keeping the head small to emphasize that improvements come from grounding rather than model scaling.

6 Weak Supervision from Human Rationales

SportR provides human-authored rationales [9]. We convert rationales into predicate labels using a constrained predicate extractor (Appendix C). Missing predicates are treated as unknown (not negative), and extraction confidence is used for loss weighting.

Noise Handling. Weak labels are inherently noisy. We mitigate this via:

- **Confidence weighting:** BCE losses are weighted by extractor confidence c_i.
- **Consistency constraints:** Composed rules must not contradict known task labels.
- **Auditing:** Extractor reliability is estimated via cross-model agreement (Appendix D).

7 Training via GRPO and RSA

Training proceeds in two stages: supervised learning with weak predicate labels, then RL post-training with risk-aware constraints.

7.1 Supervised Objective

We optimize:

$$\mathcal{L}_{\text{pred}} = \sum_{i=1}^{K} w_i \cdot \text{BCE}(\hat{p}_i, \tilde{p}_i), \tag{3}$$

$$\mathcal{L}_{\text{task}} = \text{CE}(\hat{y}, y), \tag{4}$$

$$\mathcal{L}_{\text{cons}} = \sum_{r \in \mathcal{R}} \text{BCE}(\hat{r}(\hat{\mathbf{p}}), r), \tag{5}$$

with total loss $\mathcal{L} = \mathcal{L}_{\text{task}} + \gamma \mathcal{L}_{\text{pred}} + \delta \mathcal{L}_{\text{cons}}$, where w_i reflects confidence weighting.

7.2 GRPO Post-Training

GRPO [3] computes group-relative advantages over G sampled outputs without requiring a learned critic:

$$\hat{A}_{\text{GRPO}}(x, y_i) = \frac{R(x, y_i) - \mu}{\sigma}, \quad \mu = \text{mean}(\{R(x, y_j)\}_{j=1}^{G}),\ \sigma = \text{std}(\cdot). \tag{6}$$

We define reward R to capture correctness (Q1/Q2) and evidence-faithfulness (penalizing unsupported infraction tokens when grounding evidence is low).

7.3 Risk-aware Stepwise Alignment (RSA)

To suppress false-positive infractions, we add an RSA-derived risk penalty [13]:

$$L_{\text{RSA}} = L_{\text{GRPO}} + \lambda \cdot \text{CVaR}_{\alpha}\Big[\mathbb{1}[\text{false\_positive}] \cdot c_{\text{penalty}}\Big], \tag{7}$$

where a false positive is defined as predicting an infraction when the ground truth is "no infraction." We report false-positive reduction as a relative reduction at matched recall under a calibrated infraction threshold.

8 Experimental Evaluation

8.1 Dataset and Tasks

We evaluate on SportR [9], focusing on the SportR-Hard subset. We report:

- **Q1: Infraction Identification** (accuracy %),
- **Q2: Foul Classification** (accuracy %),
- **Q5: Visual Grounding** (IoU %).

8.2 Baselines

We compare to the Qwen-VL-7B (SFT+RL) baseline reported in SportR [9] and include representative frontier model results under a consistent evaluation harness.

8.3 Main Performance on SportR-Hard

Table 2 reports performance on the SportR-Hard subset (mean $\pm$ std over 3 seeds).

Table 2: Performance on SportR-Hard. Q1 and Q2 are Accuracy (%); Q5 is IoU (%). Baseline metrics reproduced from [9].

Model	Q1 Infraction ID	Q2 Foul Class	Q5 Grounding IoU
Qwen-VL-7B (SFT+RL)	84.19 (Baseline)	51.54 (Baseline)	9.94 (Baseline)
GPT-5.2 Pro	69.19	44.21	5.70
Gemini 3 Pro	58.79	17.54	3.67
RuleGround (Ours)	**88.42 $\pm$ 0.3**	**58.60 $\pm$ 0.4**	**14.12 $\pm$ 0.2**

Table 3: Ablation study on SportR-Hard (Q1 accuracy).

Configuration	Q1 Acc (%)	Δ vs Full
Qwen-VL-7B (Baseline)	84.19	-4.23
RuleGround (Full)	88.42	–
w/o RSA Alignment	82.15	-6.27
w/o Predicate Supervision	79.50	-8.92
w/o Differentiable Logic	84.30	-4.12

8.4 Ablation Study

Table 3 evaluates the contribution of each component. The large drop without RSA indicates that tail-risk suppression directly affects infraction decisions by discouraging unsupported predictions.

8.5 Cross-Sport Transfer

Table 4 reports zero-shot performance retention for soccer→basketball rules with shared logical structures.

Table 4: Zero-shot structural transfer accuracy (soccer→basketball).

Rule Category	In-Domain (Soccer)	Zero-Shot (Basket)	Retention
Position-before-trigger	72.8	61.4	84.3%
Contact-during-play	65.4	52.1	79.6%
Mean	**69.1**	**56.8**	**82.2%**

8.6 Qualitative Examples

We provide representative examples showing how explicit predicate prediction improves adjudication.

Example 1: Basketball Blocking vs. Charging. Scenario: An offensive drive leads to a collision; adjudication depends on whether the defender is set at impact.

Table 5: Predicate comparison: Blocking/charging adjudication.

Predicate	RuleGround	Baseline MLLM (implicit)
`contact_occurred`	✓ (0.97)	✓ (mentioned)
`defender_set`	✗ (0.23)	✓ (incorrectly inferred)
`restricted_area`	✓ (0.91)	not assessed
Rule Composition	$\neg$`defender_set` $\wedge$ `contact` $\rightarrow$ **Blocking**	"Charging"
Outcome	RuleGround correct; baseline fails on `defender_set`	

Analysis: The baseline MLLM's rationale states "the defender established position before contact," but frame-level evidence shows the defender's feet were still moving at impact. RuleGround's `defender_set` classifier (trained on temporal foot-position features) correctly outputs 0.23, triggering the blocking foul rule via composition.

Example 2: Football (American) Defensive Pass Interference. **Scenario:** Receiver contacted prior to ball arrival; adjudication depends on timing and exception predicates.

Table 6: Predicate comparison: Defensive pass interference adjudication.

Predicate	RuleGround	Baseline MLLM (implicit)
`contact_occurred`	✓ (0.94)	✓ (mentioned)
`contact_before_arrival`	✓ (0.92)	✗ (claims "simultaneous")
`incidental_contact`	✗ (0.18)	✓ (incorrectly claimed)
`ball_catchable`	✓ (0.87)	not assessed
Rule Composition	`before_arrival`$\wedge\neg$`incidental`$\wedge$`catchable` $\rightarrow$ **DPI**	"No foul"
Outcome	RuleGround correct; baseline fails on timing + exception	

Analysis: The baseline misclassifies contact as "incidental" and misjudges timing as "simultaneous with ball arrival." RuleGround's temporal attention mechanism correctly estimates contact occurred 0.3s before ball arrival (`contact_before_arrival` = 0.92) and that contact materially affected the receiver (`incidental_contact` = 0.18).

9 Limitations and Discussion

RuleGround shifts the bottleneck from grounding to higher-order reasoning (Table 1). Despite gains, we observe:

Prompt Sensitivity. Performance swings up to $\sim$10% based on prompt wording in pilot runs, suggesting robustness training and paraphrase evaluation suites are important next steps.

Weak Supervision Noise. Predicate labels extracted from rationales are imperfect ($<$12% noise rate by audit); further manual annotation would strengthen grounding.

Subjective Rules. Some calls depend on officiating judgment (e.g., incidental contact thresholds), which may not be fully captured by binary predicates.

Retrieval Limitations. Dense retrieval based solely on semantic similarity can miss relational constraints central to rules; graph-structured retrieval over entities/events/predicates is a promising extension.

10 Broader Impacts

Automated adjudication should augment, not replace, human judgment. False positives can create unfair outcomes; we explicitly target tail-risk reduction via RSA. Deployment should include human-in-the-loop oversight, uncertainty communication, and auditing across leagues, camera conditions, and demographics.

11 Conclusion

We introduced RuleGround, a neuro-symbolic architecture that bridges the Perceptual Gap in sports video understanding through explicit predicate grounding. By combining a Rule Grounding Module with weak supervision from human rationales and risk-aware RL alignment, RuleGround achieves state-of-the-art performance

on SportR-Hard while providing interpretable, auditable intermediate representations. Cross-sport transfer experiments demonstrate that the architecture learns abstract rule structure rather than sport-specific patterns. Code and model weights are available at `https://github.com/sreevadde/ruleground`.

Acknowledgments

We thank the SportR benchmark authors for releasing their dataset and evaluation code.

References

[1] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pages 16344–16359, 2022.

[2] Luciano Serafini and Artur d'Avila Garcez. Logic tensor networks: Deep learning and logical reasoning from first principles to machines. In *IJCAI Workshop on Neural-Symbolic Learning and Reasoning*, 2016.

[3] Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, February 2024.

[4] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, February 2020.

[5] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, April 2021.

[6] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[7] Sree Krishna Vadde. Enhanced ActionFormer: Multi-gpu, flash attention, RoPE, snapformer, tbtformer, January 2026. Modern transformer primitives for temporal action localization.

[8] Limin Wang, Bingkun Huang, Zhiyu Zhao, et al. VideoMAE V2: Scaling video masked autoencoders with dual masking. *arXiv preprint arXiv:2303.16727*, March 2023.

[9] Haotian Xia, Haonan Ge, Junbo Zou, et al. SportR: A benchmark for multimodal large language model reasoning in sports. *arXiv preprint arXiv:2511.06499*, November 2025.

[10] Haotian Xia, Zhengbang Yang, Junbo Zou, et al. SPORTU: A comprehensive sports understanding benchmark for multimodal large language models. *arXiv preprint arXiv:2410.08474*, October 2024.

[11] Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, 2019.

[12] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision (ECCV)*, pages 492–510, 2022.

[13] Lijun Zhang et al. Constrained language model policy optimization via risk-aware stepwise alignment. *arXiv preprint arXiv:2512.24263*, December 2025.

[14] Junbo Zou et al. DeepSport: A multimodal large language model for comprehensive sports video reasoning via agentic reinforcement learning. *arXiv preprint arXiv:2511.12908*, November 2025.

A Implementation Details

Hardware. Single NVIDIA A100 80GB ($\sim$8 hours total training).

Optimizer. AdamW with learning rate 1×10^{-4}, weight decay 0.01, cosine schedule.

Hyperparameters. $\gamma = 0.5$ (predicate loss weight), $\delta = 0.1$ (consistency loss weight), $\lambda = 0.3$ (RSA penalty), $\alpha = 0.1$ (CVaR quantile), $G = 8$ (GRPO group size).

Frames. 16 frames uniformly sampled per video clip.

Temporal Attention. We use 8 attention heads with RoPE [5, 6] and Flash Attention [1] via the ActionFormer v2 implementation [7].

B Predicate Ontology

Table 7 shows the core subset of our multi-sport predicate ontology.

Table 7: Core predicate ontology (20 predicates total; representative subset shown).

Predicate	Domain	Definition	Example
`ball_in_play`	Shared	Ball is active and live	Live vs. dead ball
`contact_occurred`	Shared	Physical touch between entities	Hand-on-arm contact
`contact_before_arrival`	Shared	Contact before ball/puck arrives	Pass interference timing
`incidental_contact`	Shared	Contact not materially affecting play	Exception predicate
`offside_position`	Soccer	Beyond 2nd-to-last defender	Attacker behind line
`ball_contact_arm`	Soccer	Ball touches arm below shoulder	Handball infraction
`pivot_foot_lifted`	Basketball	Lift pivot foot before dribble	Travel initiation
`defender_set`	Basketball	Defender in legal guarding position	Charge/blocking call
`restricted_area`	Basketball	Contact in restricted area arc	Blocking exception
`ball_catchable`	Football	Thrown ball is catchable	DPI applicability

C Extraction Pipeline and Prompt

System Prompt (Predicate Extractor).

> *"You are a sports rule predicate extractor. Given a human rationale explaining a sports adjudication decision, extract all rule-relevant predicates. Output valid JSON adhering to the RuleGround schema. Only include predicates explicitly mentioned or logically implied by the rationale. Do not solve the problem, add new facts, or infer predicates not supported by the text."*

JSON Schema.

```
{
  "type": "object",
  "properties": {
    "predicates": {
      "type": "array",
      "items": {
        "type": "object",
        "properties": {
          "predicate_name": { "type": "string" },
          "value": { "type": "boolean" },
          "confidence": {
            "type": "number",
            "minimum": 0,
            "maximum": 1
          }
        },
        "required": ["predicate_name", "value"]
      }
    }
  },
  "required": ["predicates"]
}
```

D Noise / Reliability Audit Protocol

We estimate extractor reliability via cross-model agreement (Claude 4.5 Sonnet vs. GPT-5.2 Pro) on 500 samples:

- **Agreement rate:** 88.3%
- **Cohen's κ:** 0.76 (substantial agreement)

For a public release, we recommend adding manual spot-checking (e.g., 100 samples) to estimate a true label noise rate.

E Reproducibility and Deduplication Controls

We perform fine-tuning-time deduplication by ensuring SportR evaluation items (or near-duplicates under perceptual hashing and text hashing) are excluded from supervised and RL post-training data. We do not make claims about the pretraining corpora of third-party foundation models.

F Rule Composition Examples

Table 8 shows example rule compositions using differentiable logic.

Table 8: Example rule compositions in RuleGround.

Rule	Logical Formula	Sport
Blocking Foul	contact $\wedge$ $\neg$defender_set	Basketball
Charging Foul	contact $\wedge$ defender_set $\wedge$ $\neg$restricted_area	Basketball
DPI	contact_before_arrival $\wedge$ $\neg$incidental$\wedge$ball_catchable	Football
Handball	ball_contact_arm $\wedge$ $\neg$natural_position	Soccer
Offside	offside_position $\wedge$ involved_in_play	Soccer