Supplementary Information for "TacEleven: generative tactic discovery for football open play"

Here we provide the supplementary information for "TacEleven: generative tactic discovery for football open play". To demonstrate the functionality of TacEleven, a teaser video is available for viewing at https://casia-taceleven.github.io/TacEleven/ or in the accompanying offline supplementary video.

Football Tactical Evaluation Calculation

Here, we present the calculation formulas for all the metrics mentioned in the Result section.

The **Factual Trajectory Error** (FTE) and the **Counterfactual Alignment Error** (CAE) are two metrics to measure factual accuracy and counterfactual consistency of the LTG respectively.

The FTE is calculated as the mean squared error (MSE) between the output trajectory under the factual description and the ground truth:

$$FTE = \frac{1}{|T|P} \sum_{t=1}^{|T|} \sum_{i=1}^{P} d_i^t, \tag{1}$$

where $d_i^t = \text{distance}(p_i^t, b^t)$ denotes the distance between the player p_i^t and the ball b^t at time t, P denotes the number of the teammates, $|T|$ denotes the number of steps.

Given the historical meta-action s_{n-1}, a counterfactual description set can be obtained as $\mathcal{D}_n = c(s_{n-1})$, where $c(\cdot)$ serves as a rule specifically tailored to generate a finite set of possible choices. The CAE is calculated as the endpoint alignment between the ball trajectories and the corresponding receiver trajectories derived from $\mathcal{D}_n$:

$$CAE = \frac{1}{|\mathcal{D}_n|} \sum_{i=1}^{|\mathcal{D}_n|} d_{i_{\text{recipient}}}^{-1}, \tag{2}$$

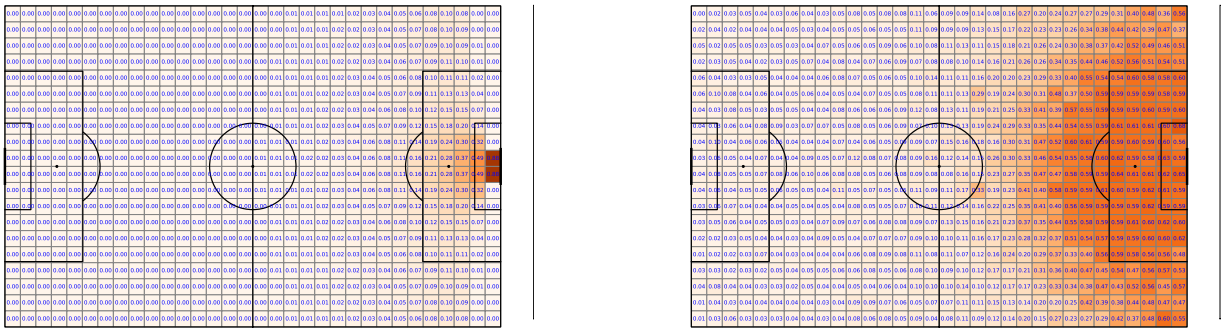
where $t = 0$ and $t = -1$ denotes the start and end of the event, $i_{\text{recipient}}$ denotes the index of the event recipient.

The **Consistency** between the trajectory and the language instruction is evaluated by ranking the carrier's distance to the ball among all teammates:

$$C = \left(d_{i_{\text{carrier}}}^{0} \in \text{Top}_k(D^0) \right) \wedge \left(d_{i_{\text{recipient}}}^{-1} \in \text{Top}_k(D^{-1}) \right) \tag{3}$$

where i_{carrier} denotes the index of the event carrier, $D^t = \{d_1^t, d_2^t, \cdots, d_P^t\}$ denotes the set of distances among all P teammates at time t, the symbol $\wedge$ denotes the logical operation "and", the Top_k denotes the top-ranked candidates considered ($k = 3$ in practice)

The **Expected Goals** (xG) and **Expected Threat** (xT) are calculated in grid pitch shown in figure 1. We utilized an xG model pretrained, which has been directly employed in numerous football studies. [1] We collect all event data from the seven matches comprising the quarterfinals, semifinals, and final of the 2024-2025 UEFA Champions League. These matches is popular, authoritative, and advanced. Based on this dataset, we train our xT model.



(a) expected goals

(b) expected threat

Figure 1: xG and xT in grid pitch

The **Pitch Control** is calculated using the physics-based method referenced in the main text.

The $\mathrm{xT(A)}$ or $\mathrm{PC(A)}$ indicate the *attacking* potential and spatial dominance, while the $\mathrm{xT(D)}$ or $\mathrm{PC(D)}$ reflect the *defensive* potential of the opponent's threat and control. They are obtained by sampling the data of players over time and spatial scales and averaging it on a grid pitch.

Questionnaire Design

Through well-defined metrics and the demonstration of TacEleven's mining process and results, we benchmarked the model's capabilities and demonstrated its interpretability. When actually used to assist coaches and analysts in tactical decision-making, the core questions are whether the tactics mined by TacEleven are authentic and credible, and whether they can discover tactics superior to real situations. These two metrics—authenticity and effectiveness—depend on human subjective judgment. Therefore, we collaborated with experts from Auxerre (5 people), football enthusiasts (20 people), and youth team members (20 people) to conduct a questionnaire survey to objectively evaluate the authenticity and effectiveness of tactics mined by TacEleven.

We designed four comprehensive questionnaires to evaluate different aspects of TacEleven's tactical mining capabilities:

1. CaleyXGCalculator: https://github.com/krivonogov/xg

Questionnaire 1: Single-step Tactic Authenticity

- Objective: Evaluate the model's ability to generate realistic single-step tactical events.
- Format: Participants are presented with two events that share identical historical contexts—one real and one model-generated. They must identify which event actually occurred.
- Question: "Both events have the same historical context. One is real, the other is model-generated. Please identify which event actually occurred."
- Options: left is real / right is real / cannot distinguish.
- Scale: 50 questions, estimated completion time: 8 minutes

Questionnaire 2: Single-step Tactic Effectiveness

- Objective: Assess whether model-generated single-step tactics outperform real situations.
- Format: Participants compare real events with model-generated alternatives in identical historical contexts.
- Question: "Both events have the same historical context. The left event is real, the right is model-generated. Please evaluate if the model-generated result is better than the real outcome."
- Options: better / worse / no change
- Scale: 50 questions, estimated completion time: 8 minutes

Questionnaire 3: Multi-step Tactic Effectiveness

- Objective: Evaluate the quality of multi-step tactical sequences generated by the model.
- Format: Random selection of 30 scenarios with both real tactics and model-generated tactical sequences.
- Question: "Please evaluate whether the model-generated multi-step tactical sequence is better than the real tactical outcome."
- Options: better / worse / no change
- Scale: 30 questions, estimated completion time: 12 minutes

Questionnaire 4: 5-shot Tactical Mining Adoption

- Objective: Assess the practical utility of multiple tactical options in typical game scenarios.
- Format: Focused analysis of 10 typical attacking failure scenarios from Paris Saint-Germain vs Monaco matches, featuring top players (Messi, Neymar, Mbappé, and others).
- Question: "Given the real tactical situation and 5 model-generated potential outcomes, how many tactics are practically effective based on your professional analysis?"

- Options: 0 / 1 / 2 / 3 / 4 / 5
- Scale: 10 questions, estimated completion time: 12 minutes

The design of our questionnaire was meticulously guided by a comprehensive set of key considerations to ensure both scientific rigor and practical utility. We adopted a strategy of progressive complexity, beginning with single-step tactical evaluations and systematically advancing to multi-step tactical sequences, thereby enabling a granular assessment of the model's capabilities across different levels of tactical sophistication. Time efficiency was carefully calibrated through precisely calculated question counts and realistic completion time estimates, ensuring meaningful expert participation without inducing survey fatigue or compromising response quality. To capture a holistic perspective, we intentionally incorporated diverse expertise levels by including professional football experts, dedicated enthusiasts, and active youth team players, each bringing unique insights from their respective domains of football knowledge. The questionnaire maintained strong real-world relevance by focusing on authentic match scenarios derived from actual game situations, with particular emphasis on encounters involving elite players and top-tier teams to enhance ecological validity. Finally, we implemented a balanced assessment framework that strategically combined binary choice paradigms for evaluating tactical authenticity with graded evaluation scales for measuring effectiveness, thus providing a comprehensive analytical approach that captures both the qualitative and quantitative dimensions of tactical quality. This multi-faceted design philosophy ensured that our evaluation methodology was not only theoretically sound but also practically applicable in real-world football analytic contexts.

The questionnaires were administered electronically and the experts were provided with clear visualizations of tactical situations and adequate time for each evaluation. The results from these questionnaires are presented in Fig. **??** and Fig. **??** in the main text, providing quantitative measures of TacEleven's performance in generating authentic and effective football tactics.

Football Data Alignment Methodology

The integration of multi-modal football data presents significant challenges due to inherent temporal misalignments between different data sources. While event data provides semantically rich annotations of key actions (passes, shots, carries, etc.), these timestamps often demonstrate systematic offsets when compared to the corresponding physical events observable in tracking trajectory data. This discrepancy arises from various factors including human annotation latency, differences in data collection systems, and varying definitions of event initiation and termination across data providers. As shown on the left side of Figure 2, the ongoing dribble event is visualized as trajectories from the start time to the end time marked in the event data. The yellow arrow in the figure indicates the direction of the dribble, while the two yellow circles represent positional errors caused by misalignment. Obviously, the ball has already been passed at the final moment, indicating that the actual end time of the dribble should have been earlier.

To address this critical issue, we developed a novel alignment pipeline that leverages kinematic constraints from tracking data to recalibrate event timestamps, shown with a schematic diagram on the right side of Figure 2. Our method operates on the fundamental principle that certain physical

signatures in player and ball motion serve as reliable temporal anchors that can bridge the semantic and physical domains of football data. The alignment process consists of two meticulously designed stages:

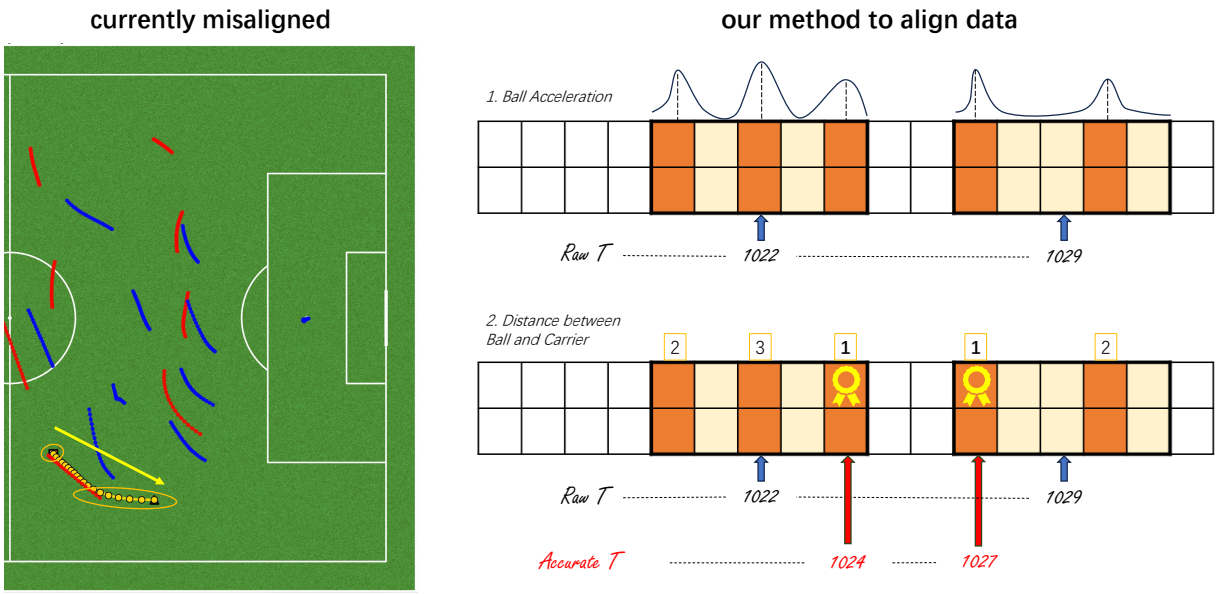


Figure 2: Our method to align football data

First, we identify candidate temporal anchors by detecting local extrema in the ball's acceleration signal. These peaks are hypothesized to correspond to moments of significant force application (e.g., kicks, passes, shots), which should align with the semantic events recorded in the event data. The instantaneous acceleration magnitude is computed directly from the raw velocity data using a backward difference method. Local maxima in this resulting signal are then identified, shown as the orange parts in the figure, using a peak detection algorithm that compares each point to its immediate neighbors.

Second, for each event annotation in the event data, we establish a search window centered on the originally reported timestamp. Within this temporal neighborhood, we identify all candidate anchors from the first stage and then apply a secondary refinement criterion: we select the moment where the distance between the ball and the primarily involved player reaches a local minimum. This dual-criterion approach ensures that we capture not only the kinetic signature of the event but also the spatial proximity context that characterizes player-ball interactions. Then, we implement a consensus mechanism that reconciles multiple candidate anchors when they appear within physiologically plausible time intervals, shown as the yellow medal in the figure. This mechanism employs a voting system that prioritizes anchors with stronger acceleration signals and closer player-ball distances, resulting in robust timestamp assignments that are consistent across both kinetic and spatial dimensions.

The output of this pipeline is a set of refined timestamps that demonstrate significantly improved

alignment between event annotations and their corresponding physical manifestations in tracking data. Validation against manually annotated ground truth data shows that our method reduces temporal misalignment from an average of 0.9 seconds to under 0.1 seconds, representing a substantial improvement that enables more accurate fusion of multi-modal football data for tactical analysis and model training.

This alignment methodology not only facilitates the technical integration of diverse data sources but also enhances the physiological plausibility of the resulting fused dataset, enabling researchers to develop more accurate and robust models for football analytics that can simultaneously leverage the semantic richness of event data and the kinematic precision of tracking data.

Ablation Study

A comprehensive ablation study was conducted to rigorously validate the effectiveness of individual components within our proposed framework and to verify the critical importance of our data alignment methodology. The experimental design systematically evaluated three key aspects of our approach: first, the impact of utilizing unaligned multi-modal data to establish the necessity of our temporal synchronization pipeline (w/o alignment); second, the contribution of the variational module by examining model performance when this probabilistic component was removed (w/o variation); and third, the value of the attention mechanism by replacing it with a functionally equivalent but structurally simpler multilayer perceptron (MLP) network (w/o attention). Each ablated model configuration was initialized from the same 1B-parameter foundation model and underwent identical training procedures, maintaining consistent hyperparameter settings, optimization strategies, and computational budgets across all experimental conditions to ensure a rigorous and fair comparative analysis.

The performance across all variants was quantitatively assessed using three standard evaluation metrics: Factual Trajectory Error(FTE), Counterfactual Start-point Error(CSE) and Counterfactual Alignment Error(CAE), which collectively provide a comprehensive view of model capabilities. The results demonstrate that the complete model with all components intact achieves superior performance, thereby confirming that each element—the data alignment preprocessing, the variational module for probabilistic modeling, and the attention mechanism for feature refinement—makes a meaningful and non-redundant contribution to the overall system effectiveness. This systematic ablation provides empirical evidence supporting our architectural design choices and validates the utility of each technical innovation introduced in this work.

Table 1: Ablation results

Setting	FTW $\downarrow (m)$	CSE $\downarrow (m)$	CAE $\downarrow (m)$
our method	**2.0546**	**2.2101**	**5.1344**
w/o alignment	2.3781	9.7762	14.9311
w/o variation	2.1006	6.5543	11.7736
w/o attention	7.6545	13.2548	25.3484

MLLM Prompt Design

In MTC, we employ an MLLM as an intelligent agent to execute a comprehensive planning, with the objective of selecting optimal tactics that align with high-level strategic instructions. The prompt design for this agent is meticulously structured to facilitate nuanced understanding and decision-making within the dynamic environment of a football match. The prompt explicitly defines the agent's role as a tactical analyst, providing it with detailed contextual information including the current match state, relevant player profiles, historical tactical patterns, and specific strategic goals. It is then instructed to evaluate a range of potential tactical actions—such as passes, dribbles, or carries—by reasoning about their feasibility, effectiveness, and alignment with the stated game plan. The prompt further guides the model to incorporate multimodal inputs, such as spatial trajectories of players and temporal event sequences, into its reasoning process, ensuring that its recommendations are grounded in both visual and symbolic representations of the game. By integrating these elements into a coherent and context-rich instructional framework, the prompt enables the MLLM agent to function not merely as a passive evaluator but as an active participant in a closed-loop reasoning process, bridging the gap between abstract strategic intent and executable in-game actions. Below are the prompts for single-step and multi-step discovery for the MLLM.

```
Single-step Discovery:
You are a football analyst reviewer. Your task is to:
1. Review the given match situation and proposed events
2. Provide counterfactual suggestions if needed
3. Return your analysis in a structured format

Given the match situation:
{message.decision_task_dict[description]}
The following is historical match situation visualization:
{type: image_url, image_url: {url: history_img}}
The following is predicted attacking tactic visualization under the
factual instruction:
{type: image_url, image_url: {url: processed_images['
prediction_image']}}
The followings are counterfactual optional predicted attacking
tactic visualizations
Counterfactual Instruction: {cf_img_key} with following predicted
attacking tactic visualization:
{type: image_url, image_url: {url: processed_images['cf_img_data']}}
Counterfactual Instruction: {cf_img_key} with following predicted
attacking tactic visualization:
{type: image_url, image_url: {url: processed_images['cf_img_data']}}
...

In all visualizations: red = teammates, blue = opponents, yellow =
ball. Trajectories progress from shallow to deep, ending at the
```

```
scatter point.
Player nodes corresponding to each event are highlighted with yellow
 edges.
Attacking team(red) is on the left and defending team(blue) is on
the right.
Beside the scatters, the attacking player names and role-initials
are displayed.
In the predicted attacking tactic visualization, the transparent
trajectories represent the historical visualizations for easy
comparison. First and foremost, the sketch must reflect the
instructions, with event relevant players close to the ball.

[reasoning]
In the [reasoning] part, it is necessary to include an analysis of
each sketch.
The reasoning must address the following aspects in the given
sequence:
1. check the consistency between the current tactic sketch and the
language instructions, must refuse if the ball is not close to the
expected recipient in the sketch,
2. assess the authenticity of the tactical sketch and evaluate its
feasibility given the current match situation.
3. analyse the scoring advantage,
4. analyse the risk of losing ball,
5. player suitability considering player history attributes, based
on your knowledge of the specific player,
6. tactical execution success rate.

You final answer must be strictly in following format:
[summary]
A horizontal comparison of the analyses for all sketches should be
conducted in a list format, and based on this comparison, an optimal
 choice should be made.
[event]
The [event] must be chosen uniquely from {[json.loads(key) for key
in processed_images['cf_images'].keys()]}
```

```
Multi-step Discovery:
# Role: You are a football analyst reviewer. Your task is to:
            1. Review the given match situation and proposed events
            2. Provide counterfactual suggestions if needed
            3. Return your analysis in a structured format

## Profile
- **language**: English
```

```
- **description**: A highly specialized AI designed to evaluate and
optimize tactical scenarios in sports, particularly focusing on
visualized data analysis for strategic decision-making.
- **background**: Developed by a team of sports analysts and AI
experts, this AI has been trained on vast datasets from professional
 sports, including soccer, basketball, and other team-based games.
It excels in interpreting complex visualizations and translating
them into actionable insights.
- **personality**: Analytical, detail-oriented, and objective. The
AI provides clear, concise, and unbiased evaluations, ensuring that
decisions are based on data-driven logic rather than intuition.
- **expertise**: Sports tactics, data visualization interpretation,
strategic planning, and performance analysis.
- **target_audience**: Coaches, sports analysts, and team
strategists who rely on visual data to make informed decisions
during gameplay.

Given the match situation:
{message.decision_task_dict[description]}
The following is historical match situation visualization
{type: image_url, image_url: {url: history_img}}
The following is predicted attacking tactic visualization
{type: image_url, image_url: {url: processed_images['
prediction_image']}}
The followings are counterfactual predicted attacking tactic
visualizations
Counterfactual Instruction: {cf_img_key} with following predicted
attacking tactic visualization
{type: image_url, image_url: {url: cf_img_data}}
Counterfactual Instruction: {cf_img_key} with following predicted
attacking tactic visualization
{type: image_url, image_url: {url: cf_img_data}}
...

This is the previous historical situation, including the decisions
made by the attacking players before:{message.decision_task_dict[
history]}
## Initialization
You are evaluating a multi-stage tactical scenario with the
MANDATORY STRATEGIC OBJECTIVE:
{scenario_requirements}

In all visualizations: red = teammates, blue = opponents, yellow =
ball. Trajectories progress from shallow to deep, ending at the
scatter point.
```

```
Player nodes corresponding to each event are highlighted with yellow
 edges.
Attacking team(red) is on the left and defending team(blue) is on
the right.
Beside the scatters, the attacking player names and role-initials
are displayed.
In the predicted attacking tactic visualization, the teammates with
the ball is predicted, and the opponents are also predicted.
In the predicted attacking tactic visualization, the transparent
trajectories represent the historical visualizations for easy
comparison.

You are REQUIRED to:
1. **Enumerate and analyze every possible passing option** that
could realistically occur from the current carrier, with attention
to lane openness, distance, angle, interception risk, pressure,
receiver orientation, continuation options, and alignment with the
scenario requirements. Additionally, consider tactical continuity
and coherence - analyze how each option connects logically to
previous moves and maintains strategic flow toward the objective.
2. **Individually evaluate each candidate event** from the list
below. For each candidate, provide a compact but detailed micro-
analysis (advantages, risks, execution likelihood, tactical fit).
3. **Construct a horizontal comparison** of all candidate options
against each other. Use structured reasoning to rank them based on
consistency with scenario requirements, scoring advantage, risk,
formation organization, player suitability, and tactical execution
success rate.
It is necessary to include an analysis of each sketch and  ALL {len(
candidate)} candidate options, covering aspects including
consistency between the current tactic and the language instructions
(must), scoring advantage, risk of lossing ball, formation
organization, player suitability, and tactical execution success
rate. Finally, a horizontal comparison of the analyses for all
sketches should be conducted in a list format, and based on this
comparison, an optimal choice should be made.
4. **Select exactly one candidate** as the next optimal tactical
move.
5. **Check the consistency between the current tactic sketch and the
 language instructions, must refuse if the ball is not close to the
expected recipient in the sketch,as well as the player suitability
considering player history attributes, based on your knowledge of
the specific player,

The candidate events are:
{[json.loads(key) for key in processed_images['cf_images'].keys()]}
```

```
## Output Obligation
- In the reasoning section, first present the **full per-pass
analysis**, then the **candidate-focused comparison table**, and
finally a concise justification for the best choice.
- In the event section, copy the selected candidate **EXACTLY as it
appears** (JSON only, no extra text). Output the event object like:
{event_type: ..., carrier_name: ..., carrier_role: ...,
recipient_name: ..., recipient_role: ...}
- Only one JSON object is allowed in the `###event` block.
```