Appendix

1 Data

Instructions for video clip caption. The list of instructions used to instruct the model to describe or caption the soccer activity shown in the video clip. They present the same meaning with natural language variance. All samples from Stage's Concept Alignment's training set use one of the following instructions.

- "Please provide a description of what happened in the soccer match video?"
- "What key events took place in this soccer video clip?"
- "Describe breifly what happened in this soccer game video?"
- "Provide a short description of the video."
- "Give an account of the activites captured in the video."
- "What is a quick summary of the soccer events in the video?"
- "Present a compact description of the video clip."
- "Relay a brief, clear account of the soccer video clip shown."

Table 5: The list of instructions for video clip description.

Instruction for Soccer Action Classification Task

```
Identify the key action that happened in the given video clip of a soccer game.

Select one from the list below:
- Ball out of play
- Clearance
- Corner
- Direct free-kick
- Foul
- Goal
- Indirect free-kick
- Offside
- Shots off target
- Shots on target
- Substitution
- Throw-in
- Yellow card

Note: Only answer with one action from the list and nothing else.
```

Figure 7: The same instruction is applied to all samples in the training and test set for the Soccer Action Classification Task. The order of the list of actions is randomized for each sample.

2 Prompt

Prompt to Generate Synthetic Captions

```
# Character
You are an AI assistant specialized in soccer topics and in anlayzing broadcast soccer
games. You are provided with 8 sequential frames of a 2 seconds long broadcast soccer
match video clip that represents one soccer event. You are provided with the
groundtruth annotation of the soccer event that took place during the clip. Using the
frames and the key soccer action groundtruth annotation, please describe or caption
what you see in the soccer match clip. Focus response on the keep soccer event of the
groundtruth label, and only add visual information when appropriate.

# Requirements and Helpful Info
Below are requirements for generating the description of the video clip:
    - Respond as if you have watched the clip instead of just select frames.
    - DO NOT make things up. Base your response only on the video frames and the
    groundtruth annotation.
    - Anonymize the team names and refer to them as team A and B, but stay consistent
    in the description. DO NOT mention the league name as well.
    - DO NOT mention jersey number and player name.
    - DO NOT mention jersey color.
    - You MUST use each frame's groundtruth annotation in your response but answer
    as if you were only given the broadcast soccer match frames.
    - DO NOT use phrases like "frame", "mentioned", "caption", "context" in the
    response.
    - Use the visual information in the video frame to make sense of the groundtruth
    annotations.
    - DO NOT start the response using 'In this broadcast soccer match' or similar.
    Answer directly with the actual description.
    - Keep your response concise.
    - Focus response on the keep soccer event of the groundtruth label.

# Input Format
Broadcast Soccer Match Frame 1: <frame1>
Broadcast Soccer Match Frame 2: <frame2>
Broadcast Soccer Match Frame 3: <frame3>
Broadcast Soccer Match Frame 4: <frame4>
Broadcast Soccer Match Frame 5: <frame5>
Broadcast Soccer Match Frame 6: <frame6>
Broadcast Soccer Match Frame 7: <frame7>
Broadcast Soccer Match Frame 8: <frame8>

Key Soccer Event in the Clip: <event>

# Output Format
Put your response in the <clip_description></clip_description> tag.
```

Figure 8: We used the above to prompt Claude 3.5 Sonnet to generate synthetic captions for the 2-second clips, which are then put in instruction following format for training in Concept Alignment.

Prompt to Generate Synthetic Question Answer Pairs

```
# Task
Given a detailed description and groundtruth soccer event that summarize the content
of a soccer game video clip, generate question-answer pairs. The question-answer
pairs should be faithful to the content of the video description and developed from
different dimensions to promote comprehensive understanding of the video. Here are
some question dimensions and their explanations and exampled question-answer pairs
for reference:
{question_definitions}

# Guidelines For Question-Answer Pairs Generation:
- Read the video description provided carefully, paying attention to the content,
such as the scene where the video takes place, the main characters and their
behaviors, and the development of the key events.
- Generate appropriate question-answer pairs based on the description. The
question-answer pairs should cover as many question dimensions and not deviate
from the content of the video description.
- Generate one question-answer pair for each question dimension.
- Multiple choice question should include multiple answer options as part of the
question. The answer must contain just one answer choice.
- Generate answer to question as if you have watched the clip and not the description.
- Generate question-answer pairs in such as a way that the answer cannot be derived
based on common sense from the question and requires understanding of the video.
- Generate most of the questions as free form answers but the rest as multiple choice
questions as shown in the examples. 4 questions free form, 1 question multiple choice.
Only 'Description' questions must have free-form answer, and other types of questions
could all be multiple choice questions.
- Team A and B are anonymized and interchangeable so don't make questions and answers
specific to team references. Make it generic.
- Use different multiple choice question and answer formats as shown in the examples.
For example, use 'A, B,...' to reference answer choices and sometimes use just '-'
to separate the choices. Also for example, sometimes the answer should just be the
answer choice, but sometimes use a full sentence that mentions the answer choice.
- Aside 'Prediction' types questions, all other question types should base their
questions and answers off of the description and ground truth soccer event(s).

# Input Format
Description: {caption},
Key Soccer Event: {event}

# Output Format:
1. Your output should be formed in a JSON file.
2. Only provide the Python dictionary string.
Your response should look like and no other text:
[{"Dimension": <dimension-1>, "Question": <question-1>, "Answer":<answer-1>,
"Is_multiple_choices": <True_or_False>},
{"Dimension": <dimension-2>, "Question": <question-2>, "Answer":<answer-2>,
"Is_multiple_choices": <True_or_False>},...]
```

Figure 9: We used the above to prompt Claude 3.5 Sonnet to generate synthetic question-answer pairs for the different question types for training in Instruction Tuning.

Question Type Definitions and Few-Shot Examples

```
# Task 1 - Description: this task is designed to assess the ability of the model in generating
an informative description of the video clip. It would essentially be rephrasing the provided
caption.
## caption-1: A player from team A, wearing the number 11 jersey, is being substituted off the
field. He hugs and celebrates with his teammates and coaching staff on the sidelines before
exiting the pitch. The player appears joyful and is all smiles as he embraces his colleagues,
likely after a positive contribution to the match.
## quesiton-1: Please provide a description of what happened in the soccer match video?
## answer-1: A player from Team A, sporting the number 11 jersey, is leaving the field as a
substitute. He shares hugs and celebrations with his teammates and coaches on the sidelines
before stepping off the pitch. The player looks delighted, beaming as he greets his colleagues,
likely in recognition of a strong performance during the game.
...

# Task 2 - Temporal: this task is designed to assess the model's capability of reasoning the
temporal order of events in the video clip, for example what activity happened before another.
## caption-1: The goalkeeper of team A takes an indirect free kick from inside the penalty area.
He kicks the ball forward towards the center of the field as the match continues.
## question-1: What happened right before the goalkeeper kicked the ball towards the center of
the field?
## answer-1: The goalkeeper took an indirect free kick from inside the penalty area, then
kicked the ball forward as the match continued.
...

# Task 3 - Causal: this task is designed to assess the model's ability to understand
the causality between events.
## caption-1: Team A takes a corner kick. The ball is played into the penalty area towards a
group of players from both teams jostling for position. Team B manages to clear the ball away
from danger as the corner kick fails to produce a scoring opportunity.
## question-1: How did the ball end up in the penalty area?
## answer-1: The ball ended up in the penalty area because Team A took a corner kick and aimed
it toward the group of players near the goal.
...

# Task 4 - Prediction: this task is designed to assess the model's ability to understand what's
going in the video clip and based of it make the right prediction of what might happen next.
## caption-1: Team A player commits a foul on a team B player in the middle of the field.
The referee blows the whistle to stop play and award a free kick to team B.
## question-1: What is likely to happen next after the recent interaction between the players
in the clip?
## answer-1: The referee will likely stop play and award a free kick to the opposing team,
Team B, because the clip shows a player from Team A committing a foul on a Team B player in
the middle of the field. According to soccer rules, a foul in open play typically results in
the opposing team gaining a free kick, allowing them to restart play from the spot of the
infraction.
...

# Task 5 - Action Recognition: this task is designed to assess the model's ability to identify
the key soccer action(s) or event(s) (such as Goal, Foul, etc) that took place in the short
video clip.
## caption-1: Team A takes a corner kick. The ball is played into the penalty area towards a
group of players from both teams jostling for position. Team B manages to clear the ball away
from danger as the corner kick fails to produce a scoring opportunity.
## question-1: What key soccer event or activity took place in the video clip?
## answer-1: The key soccer event shown in the clip is a corner kick where Team A sent the ball
into the penalty area where players from both teams were competing for position.
...
```

Figure 10: Definition of the five question types and some few-shot examples embedded in Figure 3's prompt to generate synthetic question answer pairs.

3 Additional Main Experiments Results

Metric	Claude 3.5 Sonnet	LLaMA 3.2	Base Model	3K Model	10K Model	20K Model
BLEU 1	0.154	**0.312**	0.289	0.287	0.290	0.288
BLEU 4	0.019	0.029	0.026	**0.088**	**0.088**	0.082
ROUGE	0.158	0.189	0.208	**0.264**	**0.264**	0.260

Table 6: Traditional metrics results for Caption Generation task across models.

Metric	LLaMA 3.2	Claude 3.5 Sonnet	Base Model	3K Model	10K Model	20K Model
Accuracy	0.242	0.267	0.118	0.578	0.623	**0.635**
Macro Precision	0.268	0.283	0.114	0.589	0.628	**0.644**
Macro Recall	0.245	0.266	0.118	0.578	0.623	**0.635**
Macro F1	0.199	0.253	0.092	0.578	0.623	**0.637**
Weighted Precision	0.238	0.283	0.114	0.589	0.628	**0.644**
Weighted Recall	0.242	0.267	0.118	0.578	0.623	**0.635**
Weighted F1	0.172	0.253	0.092	0.579	0.623	**0.637**

Table 7: Performance comparison for the Soccer Action Classification task across different models.

4 Ablation Studies Results

4.1 Training the Projector or Not

	Correctness Scores			Detailness Scores		
	Avg.	**SoccerNet**	**WyScout**	**Avg.**	**SoccerNet**	**WyScout**
(Question Count)	(200)	(100)	(100)	(200)	(100)	(100)
Adapted Models						
3K - LLM Only	1.76	2.11	1.40	1.955	2.29	1.62
3K - Projector+LLM	**1.92**	**2.43**	**1.41**	**2.155**	**2.55**	**1.76**

Table 8: Results comparison for the Caption Generation task between the two model variants in this ablation study

	Overall	**Question Sources**	
		SoccerNet	**WyScout**
(Question Count)	(200)	(100)	(100)
Adapted Models			
3K - LLM Only	71.85	79.356	66.35
3K - Projector+LLM	**72.99**	**80.89**	**65.09**

Table 9: Results comparison for the Visual QA task between the two model variants in this ablation study

4.2 Impact of LoRA Rank Sizes

We investigated the impact of different LoRA rank sizes by training models using Rank 32 and Rank 64. The results show very little differences and mixed results between the Caption Generation and VQA tasks. Given this, we only opted for Rank 64 when scaling up training to the 20k datasets, as it provides greater capacity for handling complex patterns.

	Correctness Scores			**Detailness Scores**		
	Avg.	**SoccerNet**	**WyScout**	**Avg.**	**SoccerNet**	**WyScout**
(Question Count)	(200)	(100)	(100)	(200)	(100)	(100)
Adapted Models						
10K - 32 Rank	2.08	2.63	1.53	2.38	2.84	1.91
10K - 64 Rank	**2.16**	**2.67**	**1.64**	**2.48**	**2.91**	**2.04**

Table 10: Results comparison for the Caption Generation task between the two model variants in this ablation study.

	Overall	**Question Sources**	
		SoccerNet	**WyScout**
(Question Count)	(200)	(100)	(100)
Adapted Models			
10K - 32 Rank	**78.15**	86.59	**69.22**
10K - 64 Rank	77.94	**87.36**	68.53

Table 11: Results comparison for the Visual QA task between the two model variants in this ablation study.

4.3 Effect of Incorporating Images in Training for Concept Alignment

Another ablation study examined whether incorporating image data alongside video data in the Concept Alignment training stage improves performance or not. While including images and training sequentially with images first then video clips produced slight gains in performance across tasks, when we increased the quantity of image data or the number of epochs it led to incoherent outputs. As a result, we excluded image data from the main experiments. However, testing training with both images and videos data may be a interesting direction for future researches to improve model performance.

	Correctness Scores			**Detailness Scores**		
	Avg.	**SoccerNet**	**WyScout**	**Avg.**	**SoccerNet**	**WyScout**
(Question Count)	(200)	(100)	(100)	(200)	(100)	(100)
Adapted Models						
10K - No Image	1.9	2.29	1.51	2.22	2.56	1.87
10K - Image then Video	**2.08**	**2.63**	**1.53**	**2.38**	**2.84**	**1.91**
10K - image and Video Mixed	1.85	2.11	1.58	2.19	2.48	1.89

Table 12: Results comparison for the Caption Generation task between the three model variants in this ablation study.

	Overall	**Question Sources**	
		SoccerNet	**WyScout**
(Question Count)	(200)	(100)	(100)
Adapted Models			
10K - No Image	78.15	86.59	69.22
10K - Image then Video	**79.04**	87.30	**70.78**
10K - image and Video Mixed	76.18	**87.48**	64.88

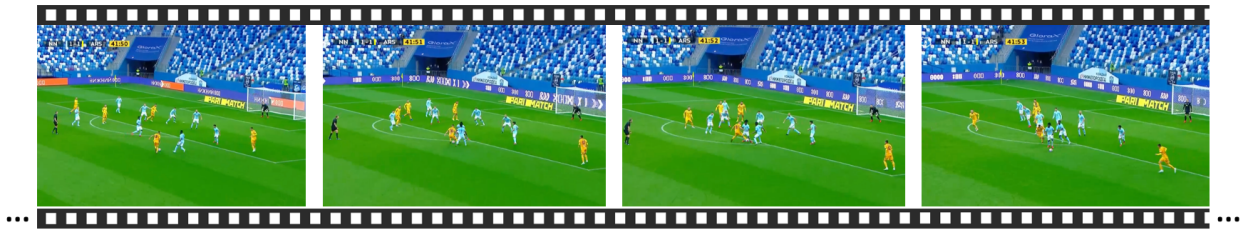
Table 13: Results comparison for the Visual QA task between the three model variants in this ablation study.

4.4 The Necessity of the Multi-Stage Training Process

Action Classification (AC)	**Accuracy**
Base Model	11.8%
Base→AC	16%
20k - Base→Concept Alignment(CA)	23%
20k - Base→CA→AC	52%
20k - Base→CA→IT→AC	**63.5%**

Table 14: Results comparison for the Action Classification task between the five model variants in this ablation study. It illustrates the values of the different training stages as the model variant that undergoes the full process achieves the best result.

4.5 Some More Qualitative Examples



Question: Which of the following key soccer events occurred in the video clip?
A. Goal scored B. Corner kick C. Defensive duel D. Penalty kick

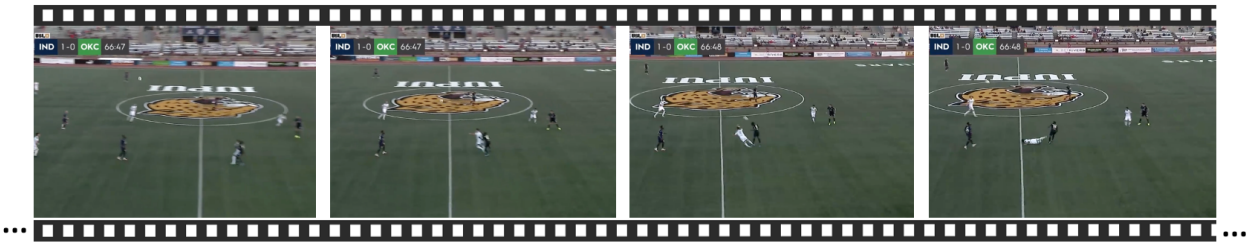
GT: C. Defensive duel

Claude 3.5 Sonnet: C. Defensive duel

Base model: The video clip shows a soccer match in progress with players on the field. There is no specific mention of a goal being scored, corner kick, defensive duel, or penalty kick in the provided information...

Adapted model: C. Defensive duel

Figure 11: A multiple choice example from the VQA task test set showing the ground-truth answer and responses from the different models. The red indicates hallucination.



Question: Can you describe the key events happening in this soccer match video?

GT: In the 67th minute of the match, with team A leading 1-0, a loose ball duel occurs near the center of the pitch... As They challenge for the ball, a player from team A commits a foul on an opponent from team B.

Claude 3.5 Sonnet: During the 67th minute of the soccer match, there was some sort of disruption or unconventional occurrence on the field. The pitch contained a large, circular graphic that seems out of place for a typical soccer game.

Base model: In the 67th minute of the soccer match, a player from the #### (wearing a white jersey) kicked the ball towards the goal, but the goalkeeper from the opposing team (wearing a blue jersey) blocked the shot...
- predicted team or player name

Adapted model: In the 67th minute, with one team leading 1-0, a foul occurred near the center circle. A player from the trailing team was in possession of the ball when a player from the leading team challenged for it, resulting in contact between the two players.

Figure 12: A free response example from the VQA task test set showing the ground-truth answer and responses from the different models. The red indicates hallucination.