Learning Consistent Temporal Grounding between Related Tasks in Sports Coaching

Arushi Rai and Adriana Kovashka

University of Pittsburgh, Pittsburgh PA 15260, USA
{arr159@pitt.edu, kovashka@cs.pitt.edu}

Abstract. Video-LLMs often attend to irrelevant frames, which is especially detrimental for sports coaching tasks requiring precise temporal localization. Yet obtaining frame-level supervision is challenging: expensive to collect from humans and unreliable from other models. We improve temporal grounding without additional annotations during training by exploiting the observation that related tasks, such as generation and verification, must attend to the same frames. We enforce this via a self-consistency objective over select visual attention maps of tightly-related tasks. Using VidDiffBench, which provides ground-truth keyframe annotations, we first validate that attention misallocation is a meaningful bottleneck. We then show that training with our objective yields gains of +3.0%, +14.1% accuracy and +0.9 BERTScore over supervised finetuning across three sports coaching tasks: ExAct, FitnessQA, and ExpertAF.

Keywords: Sports Understanding · Temporal Grounding, Localization

1 Introduction

Despite rapid advances in video understanding, video question answering models remain prone to taking shortcuts: exploiting language priors or contextual cues rather than the visual evidence that directly supports the answer [37]. This tendency is particularly problematic for sports coaching applications, where precise frame localization is essential to identify technique errors, assess form, and describe athletic actions accurately. When models bypass grounding in favor of shortcuts, users cannot trust that the feedback they receive reflects what actually occurred. Consider the frames in Fig. 1; a poorly-grounded model asked to provide technique feedback on a climber mid-route would rely mostly on language priors and contextual cues, extrapolating only the information that a person is climbing from the video. Such a model would produce common, plausible-sounding but often vague advice such as *"Pull your hips closer to the wall"*, due to limited attention to the frames that actually reveal the problem. The true technique errors are visible only in the orange frames: the climber starts pulling up, without using proper foot positioning, and straining their arms under their body weight. A well-grounded model would attend to precisely these moments before generating feedback.

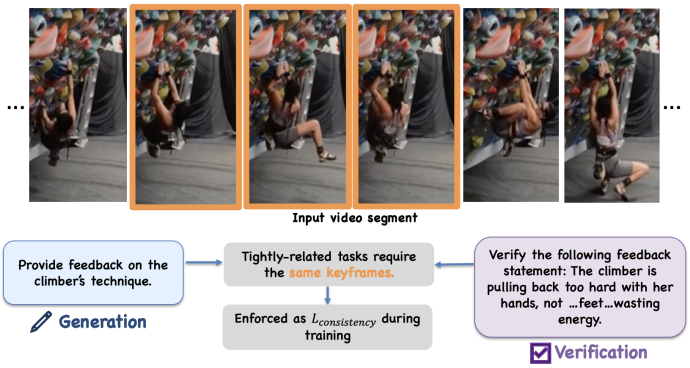


Fig. 1: Both the generation task (providing feedback on technique) and the verification task (confirming a given feedback statement) require the model to attend to the same keyframes (highlighted in orange) to produce a correct response. We exploit this as a self-supervision signal, enforcing attentional consistency between the two complementary task-views without requiring any frame-level annotations.

Two approaches to tackle the challenge have been explored in the sports domain. The first collects real frame-level annotations [7], but this is prohibitively expensive due to the large number of frames in a typical video, and is limited to a single benchmark, not a training dataset. The second, explored in concurrent work, uses pseudo-labels obtained from other models [46], but risks propagating hallucinations about which frames actually matter. This motivates a self-supervised approach that encourages the model to attend to the right temporal evidence without requiring any additional frame-level annotations.

Self-supervised techniques in computer vision have successfully learned visual representations by exploiting invariances across different views of the same image, such as augmented crops, color jitter, or temporal frames, without requiring labels [12,18,19]. Extending this idea to video-language models is non-trivial, as the input space is multimodal and the meaningful notion of a "view" is no longer obvious. One natural approach is to treat task prompt rephrasings as augmented views, encouraging the model to produce consistent representations across semantically equivalent inputs. However, rephrasings (e.g. "What feedback would you give this athlete?", "Is this athlete's form correct? Yes or No", or "Summarize this video") may fundamentally change the nature of the expected answer, and so enforcing hidden state or answer consistency is ill-defined. Additionally, even if answers or hidden states are consistent, the model may attend to entirely different and potentially irrelevant frames in each rephrased task. Hence, output-level consistency does not guarantee temporal consistency.

We address the challenge of learning to attend to the right frames, by enforcing consistency at the level of visual attention maps, and using tightly-related tasks. We observe that tightly-related tasks, such as generating technique feedback and verifying whether a given feedback statement is correct, require the

model to attend to the same keyframes regardless of differences in output form (shown in Fig.1). While we do not know which frames these are, we use the observation about their *sameness* as a self-supervised loss to guide the model's attention. We choose verification as guidance because it is more concrete (verify specific feedback claim) than the generation task where vague feedback and inaccurate attention might not be penalized.

Concretely, we route the same video through a shared model under two task prompts (i.e., a feedback generation pathway and a verification pathway). In *select vision-centric* layers and attention heads, we train the generation pathway to match the attention maps of the (simpler) verification pathway using a self-consistency loss. This provides a direct, annotation-free training signal that encourages the model to localize the same temporal evidence regardless of the exact form of the task it is performing.

To identify which layers and attention heads are most relevant for visual localization, we first conduct an analysis on VidDiffBench [7] (a video difference benchmark with frame-level annotations) in Sec. 4.1. Using this dataset, we characterize visual attention quality across layers and heads, and validate that manipulating attention toward ground-truth keyframes (as proof-of-concept) improves performance. We then apply our self-consistency method, which **requires no frame-level annotations at training time**, and evaluate across a suite of sports coaching benchmarks: ExAct [40] (multiple-choice QA that evaluates expert-level understanding of physical skills), FitnessQA [27] (fine-grained fitness coaching QA), and ExpertAF [3] (expert-level sports feedback generation).

Our method consistently improves over the finetuned baseline across all sports coaching benchmarks, showing that self-consistency over select vision-centric heads yields meaningful gains orthogonal to those from supervised finetuning alone, even though generation and verification need not attend to identical frames in every instance.

To summarize, our contributions are:

- An analysis of visual attention quality across layers and heads in a video-language model, identifying attention misallocation as an independent bottleneck for sports understanding.
- A dual-pathway self-consistency training framework enforcing attentional agreement between complementary task views without frame annotations during training.
- Significant improvement (gains of +3.0% and +14.1% accuracy on ExAct and FitnessQA, and +0.9 BERTScore on ExpertAF) over supervised finetuning in performance across multiple sports understanding benchmarks.

2 Related Works

Temporal grounding and localization. A significant body of work improves temporal grounding in Vid-LLMs through architectural modifications [26, 29], chain-of-thought reasoning over frames [1, 46] or iteratively refining input [14,

22, 33]. While effective, these approaches require architectural changes, additional supervision, or increased inference cost. Weakly-supervised localization works [4, 5, 21, 30] also exploit consistency but differ from ours in terms of architectures and consistency formulation. [5, 21, 30] do not target Vid-LLMs, where the absence of modality-specific modules makes consistency placement across attention heads non-trivial (Sec 4.4); [4] applies consistency over paraphrased outputs and [5] across cross-modal feature representations, whereas we exploit *attention* consistency across related tasks.

Targeting attention distributions. A complementary direction targets the internal attention distributions of existing models directly, without modifying the architecture or requiring additional annotations. Attention exists as a mechanism for capturing relationships between tokens in transformer models [32], making it a natural lens for explainability [9] and, more recently, for diagnosing and correcting failures in multimodal LLMs. [11] connects the persistent challenge of spatial reasoning in MLLMs to failures in visual attention, finding that uniformly-distributed attention is preferable to confidently concentrated but incorrect attention, and proposes a confidence-based intervention at inference time. [23] similarly operates post-hoc, identifying image-centric attention heads and redistributing weight away from visual sink tokens that accumulate high attention without contributing meaningfully to the output. Post-hoc attention steering has also been explored via user-specified emphasis, both in language models [42] and extended to the multimodal setting [44]. A related line of work uses attention to guide token pruning, retaining only the most relevant frames or tokens for downstream reasoning [22]. Closer to our approach, [24] performs attention distillation from a larger teacher model to a smaller student, finding that attention misalignment is a significant bottleneck in transferring visual perception. Our method differs from all of the above in that it does not rely on post-hoc steering (although we test this in our analysis in Sec. 4.1), external supervision, or a separate additional teacher model. Instead, we enforce visual attention consistency between two task formulations of the same model, exploiting task invariance of visual attention as a self-supervised training signal.

Sports understanding. General-purpose MLLMs struggle on sports understanding tasks [3, 7, 15, 20, 27, 28, 36], in part due to the domain knowledge gap and spatio-temporal demands of the task. Several works have sought to close this gap by incorporating richer domain supervision or domain-knowledge graphs [10, 15, 28, 35, 39]. [3] introduces the task of generating actionable expert feedback from video, jointly training on feedback generation, expert video retrieval, and 3D pose generation. [2] proposes to jointly train feedback generation and classify transferable skill-attributes such as balance and hand positioning that generalize across sports. More relevant to our work, recent **concurrent** efforts have explored post-training approaches to improve sports reasoning and temporal localization. [6] shows that reinforcement learning post-training alone does not necessarily improve performance on tasks requiring temporal grounding when sport-specific domain knowledge is not needed, suggesting that RL primarily enhances general content understanding rather than temporal ground-

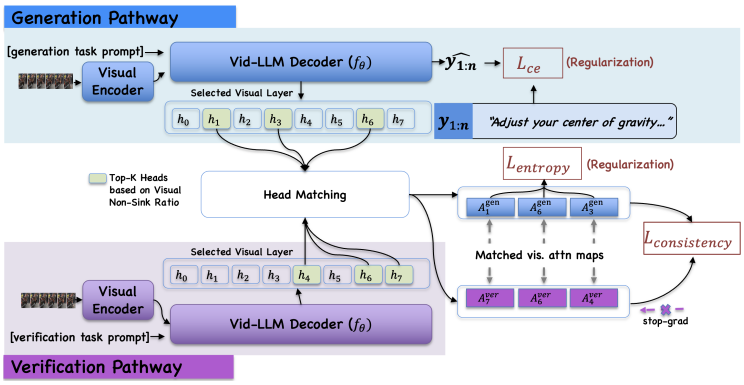


Fig. 2: Overview of the proposed dual-pathway self-consistency framework. The **Generation Pathway** and **Verification Pathway** process the same video through a shared visual encoder and Vid-LLM Decoder (f_θ) with different task prompts. From a selected visual layer (Sec. 4.1), top-K attention heads are identified via their Visual Non-Sink Ratio [23] and paired across pathways via **Head Bipartite Matching** (exact index matching + Hungarian algorithm). The generation pathway's attention maps are regularized by $\mathcal{L}_{entropy}$ for focused attention over specific visual tokens (and frames), while $\mathcal{L}_{consistency}$ enforces cross-pathway attentional agreement between matched heads. $\mathcal{L}_{ce}$ serves as regularization on the generation pathway output.

ing. This is further supported by DeepSport [46], a sports-specific reasoning model that uses chain-of-thought traces from open-source models [38] for explicit temporal localization, yet has limited improvement on assessment and coaching questions compared to recognition tasks (rule application, foul detection, action recognition) despite localization-specific RL post-training. DeepSport's error analysis of the trained model reveals frequent hallucination of temporal timestamps on fine-grained tasks, suggesting that post-training alone or distilled traces are insufficient to ground the model's attention in the correct frames. SportR [35] similarly finds through error analysis that visual perception and hallucination comprise the majority of errors from their sports reasoning model. Motivated by this, our method targets temporal grounding directly through self-supervised attention consistency, without relying on any chain-of-thought supervision or frame-level annotations, and shows consistent gains on assessment and coaching tasks [3, 27, 40]. Our method could be considered complementary to chain-of-thought supervision and post-training.

3 Method

We propose a dual-pathway self-consistency framework for video-language model (Vid-LLM) training that improves internal visual localization in sports coaching tasks **without external grounding supervision**. Self-attention in middle LLM decoder layers localizes task-relevant visual tokens [45], making them key

to how the model grounds its responses. We observe that these select attention modules should attend to consistent visual regions for two tightly related tasks, and exploit this by enforcing attentional consistency between select vision-centric attention maps from two task-conditioned pathways: a **generation pathway** (harder task) and a **verification pathway** (easier task). Figure 2 provides an overview. If the model is genuinely grounded, both tasks should localize the same visual regions, and encouraging this agreement provides a useful inductive bias that discourages spurious or task-inconsistent visual attention.

3.1 Preliminaries

Let $x^{vis} = \{x_1^{vis}, \ldots, x_T^{vis}\}$ denote a sequence of T sampled video frames. A Vid-LLM processes $\mathcal{V}$ via a visual encoder f_ϕ to produce frame-level visual tokens, which are concatenated with a text prompt x^{text} and passed to an autoregressive decoder f_θ to generate a response $\hat{y}_{1:n}$. Formally:

$$\hat{y}_{1:n} = f_\theta\big(f_\phi(x^{vis}),\, x^{text}\big). \tag{1}$$

3.2 Dual-Pathway Forward Pass

Given a video x^{vis} and a paired generation-verification prompt pair $(p^{\text{gen}}, p^{\text{ver}})$, we perform two forward passes through the shared frozen encoder f_ϕ and trainable decoder f_θ.

Generation Pathway. The generation pathway takes the video $\mathcal{V}$ and a generation task prompt p^{gen} (e.g., a sports feedback [3] or sports understanding QA instruction [27, 40]) and produces predicted output tokens $\hat{y}_{1:n}$, supervised by a standard cross-entropy loss:

$$\mathcal{L}_{ce} = -\sum_{t=1}^{n} \log f_\theta\big(y_t \mid y_{<t},\, f_\phi(x^{vis}),\, p^{\text{gen}}\big). \tag{2}$$

Within our self-consistency training framework, this loss serves as a regularization term that preserves the model's generation capability.

Verification Pathway. The verification pathway processes the same video $\mathcal{V}$ with a verification task prompt p^{ver} through the same shared decoder f_θ. The verification task prompt is constructed by combining the ground-truth answer $y_{1:n}$ with an automatically rephrased version of the generation prompt p^{gen}, forming a yes/no question that asks whether $y_{1:n}$ is a valid response to the video given the rephrased prompt (e.g. "Please verify if the following feedback statement is correct for the given video: answer"). **No gradients flow through the verification pathway**, to prevent training collapse [8, 18, 19]. The verification pathway contributes to training exclusively through the self-consistency loss $\mathcal{L}_{consistency}$ applied to its attention maps, with the stop-gradient operator (Sec. 3.5) ensuring that the verification pathway acts purely as a target signal.

3.3 Visual Layer and Head Selection

The choice of both heads and layer is critical, as not all heads and layers encode meaningful temporal information. We adopt different selection strategies for each. For the *layer*, we are motivated by prior findings in [45], which suggests that the layers most responsible for encoding task-specific visual information remain **consistent** across similar tasks. Prior to training, we identify the layer with the most temporal grounding potential, ℓ^*, using the visual attention quality score (Eq.6) which captures sharpness, vision-centricity, and keyframe overlap, each an intuitive proxy for temporal grounding. While keyframe overlap is computed using a small annotated disjoint set from the training data, the selected layer, ℓ^*, is then fixed throughout training and our method requires no additional temporal grounding labels. We discuss this in more detail and provide empirical evidence in Sec. 4.1.

Given a fixed layer, we found that *static* head selection based on the visual attention quality score is suboptimal empirically. We therefore select the top-K heads *dynamically* per forward pass using the Visual Non-Sink Ratio [23] (VNSR), which measures how much visual attention falls on non-sink versus sink tokens (see Sec. 1 of the supplement for the full definition). Since visual sink tokens attract disproportionately high attention due to large feature-dimension activations despite carrying no task-relevant content, VNSR favors heads whose attention is concentrated on meaningful visual content. We select the top-K heads from each pathway ranked by VNSR, yielding head sets $\mathcal{H}^{\text{gen}} = \{h_1^{\text{gen}}, \ldots, h_K^{\text{gen}}\}$ and $\mathcal{H}^{\text{ver}} = \{h_1^{\text{ver}}, \ldots, h_K^{\text{ver}}\}$. We use $K = 5$, and ablate in exp (Sec.4.4).

3.4 Head Bipartite Matching

To enforce self-consistency, we first must establish a correspondence between heads in $\mathcal{H}^{\text{gen}}$ and $\mathcal{H}^{\text{ver}}$. We propose **Head Bipartite Matching**, a two-stage procedure that combines exact index matching with the Hungarian algorithm [25] for assignment.

Stage 1: Exact Index Matching. We first match heads that share the same index across both pathways, i.e., $h_i^{\text{gen}} \leftrightarrow h_i^{\text{ver}}$ if $i \in \mathcal{H}^{\text{gen}} \cap \mathcal{H}^{\text{ver}}$ within the selected visual layer.

Stage 2: Hungarian Matching. For remaining unmatched heads, we compute a pairwise cost matrix $\mathbf{C} \in \mathbb{R}^{K' \times K'}$ between the unmatched heads of each pathway, where K' is the number of remaining heads. The cost between head i from the generation pathway and head j from the verification pathway is the asymmetric KL divergence $D_{\text{KL}}(\mathbf{A}_j^{\text{ver}} \| \mathbf{A}_i^{\text{gen}})$, computed over visual token indices $\mathcal{I}_{\text{vis}}$, consistent with the directional formulation of $\mathcal{L}_{\text{consistency}}$ (Eq. 4). The Hungarian algorithm then finds the optimal 1:1 assignment permutation $\sigma^* : \{1, \ldots, K'\} \to \{1, \ldots, K'\}$, and the combined matching from both stages yields a full set of K matched head pairs $\{(h_i^{\text{gen}}, h_{\sigma^*(i)}^{\text{ver}})\}_{i=1}^{K}$.

3.5 Losses

Let $\mathbf{A}_i^{\text{gen}}$ and $\mathbf{A}_i^{\text{ver}}$ denote the visual attention maps produced by the i-th matched head pair from the generation and verification pathways, respectively, where each map $\mathbf{A_i} \in \mathbb{R}^{1\times N_{\text{vis}}}$ captures the attention from the last prompt token to all visual tokens, $N_{vis} = |\mathcal{I}_{\text{vis}}|$.

Entropy Regularization. To encourage the generation pathway's attention heads to produce focused, peaky attention distributions over visual tokens rather than uniform or sink token-dominated maps, we apply an entropy regularization loss over the selected vision-centric heads:

$$\mathcal{L}_{entropy} = \frac{1}{K}\sum_{i=1}^{K}\mathcal{H}\big(\frac{\mathbf{A}_i^{\text{gen}}}{\tau_{entropy}}\big), \tag{3}$$

where $\mathcal{H}(\cdot)$ denotes the Shannon entropy computed over the visual token dimension and $\tau_{entropy}$ scales the distribution to be sharper to help convergence.

Self-Consistency Loss. The central contribution of our method is a training signal that requires no frame-level annotations during training: since both generation and verification pathways must localize the same frames, we penalize disagreement between their matched attention maps and apply a stop-gradient to the verification pathway:

$$\mathcal{L}_{consistency} = \frac{1}{K}\sum_{i=1}^{K} D_{\text{KL}}\Big(\text{sg}\big(\mathbf{A}_i^{\text{ver}}\big)\ \Big\|\ \mathbf{A}_i^{\text{gen}}\Big), \tag{4}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. The asymmetric direction treats the verification pathway's matched attention maps as a fixed reference toward which the generation pathway is trained; since verification is a simpler and more concrete task, its attention maps provide a more reliable grounding target. This design prevents training collapse, analogous to the fixed target network in BYOL [18] and MoCo [19].

Total Training Objective. The final training objective combines all three losses with equal weighting:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{entropy} + \mathcal{L}_{consistency} \tag{5}$$

We find this equal weighting to be robust across our experimental settings, but future work could add additional hyperparameters.

4 Experiments

To precisely apply our method to modules related to temporal grounding, we define heuristics and identify suitable candidate layers and heads in Sec. 4.1. We then verify these modules' sensitivity to temporal grounding quality by redistributing visual attention toward ground-truth keyframes and away from irrelevant frames in Sec. 4.2. We additionally test strategies for adaptive head

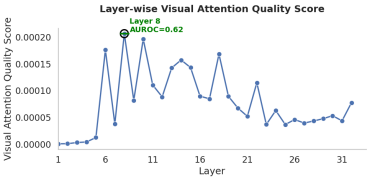
selection, which we adopt in our method. These results inform which attention modules we target with the self-consistency and entropy regularization losses. We show gains on all three datasets in Sec. 4.3 and ablate each component in Sec. 4.4.

Experiment Details. All experiments use 1 GPU with an aggregated batch size of 4, implemented via gradient accumulation over 4 steps with a per-step batch size of 1. We use a learning rate of 2×10^{-4} with LoRA (rank 16, alpha 32, dropout 0.05) applied solely to the language model. We set $\tau_{entropy}$ to 0.03, to significantly sharpen the internal visual attention due to very high initial entropy. We use PyTorch with PerceptionLM [13] as our base model, with the visual pooling rate (controls # of tokens per frame in PerceptionLM [13]) as 5 and uniformly sample 16 frames during training due to GPU memory constraints. During inference, visual pooling rate is set to 4 for a richer visual representation, and the sampling temperature is set to 1.

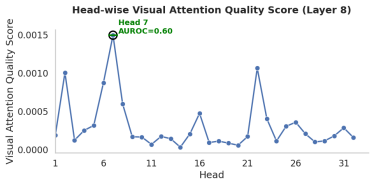
Datasets. We select three sports coaching datasets that test fine-grained understanding and feedback generation. ExAct [40] and FitnessQA [27] both evaluate fine-grained sports coaching understanding, with questions targeting technique quality, cause-effect of motions, and similar aspects. ExAct is drawn from EgoExo4D [17], covering physical skilled demonstrations across sports ranging from soccer to dance, and provides curated multiple-choice questions for fine-grained understanding. FitnessQA covers exercise demonstrations (squats, push-ups, etc.) with fine-grained questions about movement quality, but uses open-ended answers. Since open-ended answers require either an LLM-as-a-judge or imprecise semantic similarity metrics to evaluate, we use an LLM to convert the original annotations to multiple-choice questions (prompt, examples, and human evaluation results provided in Sec. 3 of the supplement), enabling straightforward accuracy-based evaluation. ExpertAF [3] is a feedback generation task, where the model must produce open-ended coaching feedback rather than answer specific questions about technique, making it unsuitable for multiple-choice conversion. We therefore evaluate predicted feedback against reference text using BERTScore [43]. We train on 800 samples each from ExAct [40], FitnessQA [27], and ExpertAF [3]. We evaluate on a 200-sample test split of ExAct [40], 659-sample test split of FitnessQA [27], and the entire test-split of ExpertAF (7K samples).

4.1 Which attention modules are sensitive to temporal grounding?

Before validating our method, we first seek to understand which attention modules selectively attend to tokens from relevant frames, and then later in Sec. 4.2 test if they can be perturbed to improve sports coaching performance. Prior work has explored the characteristics of these attention distributions within the image context of MLLMs [11, 41, 45] finding that certain layers and heads are more vision-centric than others, and that similar mid-level layers tend to specialize in visual grounding consistently across similar tasks. However, since this work only examines spatial grounding in images, it is unclear whether the same patterns hold for temporal grounding in video. To find which layer is a good



(a) Layer-wise visual attention quality score.



(b) Head-wise visual attention quality score.

Fig. 3: Visual attention quality score visualization: product of attention sharpness (1 − entropy), vision centricity (visual attention sum), and keyframe overlap (AUROC). Both graphs show that layer 8 and head 7 are relatively better at temporal grounding.

candidate for **improving** temporal grounding in the video domain, we use the EgoExo4D [17] split of VidDiffBench [7], which provides ground-truth keyframe annotations per question; these are all known to be relevant to the task.

Defining temporal grounding sensitivity. A layer may be sensitive to temporal grounding yet still attend to the wrong frames, meaning it has the capacity to ground but is currently misaligned. To identify such layers, we propose heuristics that characterize attention quality along two axes, in addition to current temporal grounding ability: (1) *selectiveness*, measured by attention sharpness since select frames should contribute to the answer in this task; (2) *vision-centricity*, which measures how much attention mass is allocated to visual tokens, compared to text or system tokens; and (3) *keyframe overlap*, measured by the AUROC between the attention distribution over visual tokens and the ground truth keyframe indicator; a similar characterization for objects within a single image was proposed in [11]. We combine these three axes into a **single visual attention quality score**:

$$S(\bar{A}^\ell) = \underbrace{\left(1 - \frac{H(\bar{A}^\ell)}{\log N_{\text{visual}}}\right)}_{\text{selectiveness}} \times \underbrace{\sum_{i \in \mathcal{V}} \bar{A}^\ell_i}_{\text{vision-centricity}} \times \underbrace{\text{AUROC}(\bar{A}^\ell{}_{\mathcal{V}}, \mathbf{k})}_{\text{keyframe overlap}} \tag{6}$$

where $\bar{A}^\ell \in \mathbb{R}^N$ is the mean attention vector at layer ℓ, averaged over all heads and samples, $\bar{A}^\ell{}_{\mathcal{V}} \in \mathbb{R}^{N_{\text{visual}}}$ is its restriction to visual token indices $\mathcal{V}$, $N_{\text{visual}} = |\mathcal{V}|$, and $\mathbf{k} \in \{0,1\}^{N_{\text{visual}}}$ is the ground truth keyframe indicator where

$$\mathbf{k}_i = \begin{cases} 1 & \text{if token } i \text{ belongs to a keyframe} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

The entropy (for computing sharpness) is calculated over the visual attention distribution and captures its peakiness:

$$H(\bar{A}^\ell) = -\sum_{i \in \mathcal{V}} p_i \log p_i, \quad p_i = \frac{\bar{A}^\ell{}_i}{\sum_{j \in \mathcal{V}} \bar{A}^\ell{}_j}. \tag{8}$$

Table 1: Proof-of-concept analysis: Effect of manipulating attention distributions using ground-truth *keyframes* (relevant frames) on the EgoExo4D split of VidDiffBench [7]. Biasing attention toward annotated keyframes improves accuracy over uniform or no redistribution, demonstrating the benefit of temporal (frame) grounding. Regarding head selection, automatic selection via visual non-sink ratio [23] performs best.

Head Selection	**Redistribution**	**Accuracy**
Manual Head Selection (layer=8, head=7)		
None	None	65.8
None	Uniform over all frames	68.4
None	Proportional over keyframes	**71.8**
Adaptive Head Selection (layer=8)		
Top-K vis. attention sum	Proportional over keyframes	69.2
Top-K vis. non-sink ratio	Proportional over keyframes	**73.4**

Layer and head manual selection analysis. As shown in Fig.3(a), layer 8 achieves the highest visual attention quality score. However, the keyframe overlap (AUROC) is 0.62, where 0.5 is no better than random guessing, which reveals that the model's internal attention does not reliably localize ground truth keyframes, indicating that the bottleneck could lie in visual representation or attention allocation. When inspecting individual heads within layer 8 in Fig.3(b), head 7 yields the highest quality score, yet we find that its average attention is heavily concentrated on the first few visual tokens regardless of where the keyframes appear. We attribute this to the visual sink token phenomenon [23], where large feature-dimension activations disproportionately attract attention mass. Since prior work finds that vision-centric layers are consistent across similar tasks [45], we fix the layer to layer 8 and use adaptive head selection (described next) to favor heads not dominated by visual sink tokens for our method.

4.2 Can temporal grounding improve performance independently of visual representation?

Our method targets attention misallocation as a meaningful bottleneck. In this section, we use ground-truth keyframe labels to test whether correcting attention misallocation improves performance, and to inform two design choices: fixing the layer and using adaptive head selection.

Attention redistribution strategies. To evaluate if reallocating the visual attention based on keyframes (relevant frames) can improve performance, we compare attention reallocation against two baselines: no manipulation, and uniform redistribution over all visual tokens (without considering keyframes). Since modern MLLMs [13, 34, 38] represent each frame with multiple tokens, we redistribute the visual attention *proportionally over keyframes* which preserves the relative within-keyframe attention distribution (see figure in Sec. 2 of the

Table 2: Performance comparison across sports understanding benchmarks. Our method improves over the finetuned baseline on all evaluated datasets. * denotes discarding answer extraction failures

Method	ExAct	FitnessQA	ExpertAF
Gemini-3-Flash	79.0	54.2	–
Qwen2.5-VL-7B-Instruct	–	33.3*	27.3
Qwen3-VL-8B-Instruct	40.2	55.4*	27.1
PerceptionLM-8B	27.1	39.5	26.0
PerceptionLM-8B (Finetuned)	84.5	74.1	37.5
+Ours	**87.5**	**88.2**	**38.4**

supplement). As shown in top half of Table 1, the proportional redistribution performs the best (71.8) compared to uniform redistribution and no manipulation. This shows potential impact for our method: if the model learns to attend better in these temporal grounding sensitive modules, performance will improve.

Adaptive head selection. We explore adaptive identification of the top-K vision-centric heads in layer 8, since the heads activated change per input (motivating the dynamic head selection in our method). We test using both the visual non-sink ratio [23] or the visual attention sum to rank and select top-K heads ($K = 5$). In the bottom half of Table 1, we observe that selecting by visual non-sink ratio is more effective (73.4) than by visual attention sum, as the latter risks including sink-dominated heads that allocate high attention mass to tokens not carrying task-relevant visual information. This informs our choice to select top-K heads using the visual non-sink ratio in our method. Still, both of these head selection methods (and even static head selection) improve over no or uniform redistribution. This confirms that a large limitation is attention misallocation rather than solely model capacity or visual representation: correcting which frames the model attends to, even post-hoc, yields consistent gains, **motivating our method to do so without requiring ground-truth keyframe labels.**

4.3 Does targeting temporal grounding sensitive modules with our method improve sports coaching performance?

We evaluate our method on three sports coaching benchmarks and compare against several baselines. The most direct comparison is the finetuned PerceptionLM-8B baseline, and we additionally compare against the larger, closed-source Gemini-3-Flash [16] and two open-source Vid-LLMs, Qwen2.5-VL-7B-Instruct [31] and Qwen3-VL-8B-Instruct [38]. We include additional results on attention quality, extra ExpertAF metrics, generalization to Qwen2.5-VL-7B-Instruct, stop-gradient ablations, and qualitative examples in the supplement.

As shown in Table 2, our method consistently improves over the fine-tuned PerceptionLM-8B baseline across all evaluated datasets. We observe gains of

Table 3: Ablation over head matching strategies and task for $\mathcal{L}_{\text{consistency}}$. "Fixed" denotes that both pathways share the same selected layer ℓ^*; "Flexible" allows the verification pathway to select its own layer independently. "Verification → Summarization" replaces the verification pathway prompt with a summarization task.

Verif. Layer	Head Matching Strategy	ExAct	FitnessQA
Fixed ($\ell^* = 8$)	Ex. Match + Hung. (ours)	**87.5**	**88.2**
Fixed ($\ell^* = 8$)	Ex. Match + Random	83.0	85.4
Fixed ($\ell^* = 8$)	Ex. Match + Discard	82.5	81.3
Flexible	Hung. only	83.0	87.7
Verif. → Summ.	Ex. Match + Hung.	69.5	83.6

Table 4: Ablation over number of heads selected.

K	ExAct
5	87.5
10	84.5
15	80.0
25	81.5
32	63.5

Table 5: Ablation over loss components.

L_{ce}	$L_{entropy}$	$L_{consistency}$	ExAct	FitnessQA	ExpertAF
✓			84.5	74.1	37.5
✓	✓		77.5	77.7	37.5
✓		✓	85.0	86.8	36.8
✓	✓	✓	**87.5**	**88.2**	**38.4**

+3.0 on ExAct, +14.1 on FitnessQA, and +0.9 BERTScore on ExpertAF, demonstrating that enforcing attentional self-consistency during training leads to meaningful improvements across tasks requiring precise localization under diverse output formats and evaluation criteria. Crucially, these gains are achieved without additional annotations during training, increased inference cost, or architecture modifications.

Notably, an 8B open-source model can surpass closed-source systems on fine-grained sports understanding. Our method achieves 87.5 on ExAct and 88.2 on FitnessQA, outperforming Gemini-3-Flash (79.0/54.2) despite substantial differences in model scale and training data. While visual representation quality prior to the LLM decoder and overall model scale remain important factors, these results suggest that attention localization constitutes an independent and addressable bottleneck in addition to domain knowledge. With model capacity and visual representations held fixed, explicitly encouraging consistent temporal attention yields further performance gains.

4.4 Ablation Studies

We conduct ablation studies to investigate the effectiveness of each component of our method. Table 3 ablates the verification layer, the strategy used to match unmatched heads in $\mathcal{L}_{\text{consistency}}$, and choice of related task. We consider three alternatives to our proposed Exact Match + Hungarian strategy

(first row): (1) **randomly** matching the unmatched head sets from each pathway, (2) **discarding** unmatched heads entirely, and (3) allowing the verification pathway to **flexibly** select its visual layer, using the top-K heads from the generation pathway as anchors for matching. Both random assignment (83.0/85.4 accuracy on ExAct/FitnessQA) and discarding (82.5/81.3) underperform our method (87.5/88.2), confirming that the quality of head correspondence directly determines the quality of the consistency signal. Discarding might be particularly harmful at $K = 5$, where losing even 1-2 unmatched heads leaves too few heads contributing to the attention-based losses, whereas random assignment at least ensures all K heads participate despite the noisy correspondence. The flexible layer strategy also underperforms (83.0/87.7) compared to our method (87.5/88.2), indicating that sharing the same layer ℓ^* across pathways is important for producing attention maps that are comparable.

We additionally ablate the choice of the related task. We replace the verification task which is tightly related to the generation task, with a general summarization task ("Summarize this video."). This leads to a significant drop in performance (69.5/83.6) compared to our method, confirming that the complementary task structure is essential. Summarization distributes attention broadly across all frames, and so the consistency signal would penalize the generation pathway for attending selectively to the relevant frames. The comparably smaller drop on FitnessQA may be partly due to the repetitive nature of fitness exercises, where relevant frames could be more broadly distributed across the video, partially overlapping with the diffuse attention of the summarization task.

Number of selected heads. Table 4 shows that $K = 5$ achieves the best performance (87.5), with accuracy sharply degrading as K increases. We attribute this to the inclusion of non-vision centric heads at larger K values, which introduce noisy attention maps that dilute the consistency signal.

Loss components. Table 5 ablates the contribution of each loss term. Removing $\mathcal{L}_{consistency}$ while keeping only the regularization losses leads to a significant drop (77.5 on ExAct / 77.7 on FitnessQA), confirming that $\mathcal{L}_{consistency}$ is essential to improving understanding capability rather than only enforcing peakiness through the $\mathcal{L}_{entropy}$. Using $\mathcal{L}_{ce}$ with $\mathcal{L}_{consistency}$ alone (row 3, 85.0 / 86.8) already surpasses the finetuned baseline, demonstrating that the consistency signal is the primary driver of improvement. Adding $\mathcal{L}_{entropy}$ (full model, row 4) further improves performance (87.5 / 88.2) and improves from 36.8 to 38.4 on ExpertAF, validating its role in encouraging focused attention distributions.

5 Conclusion

Sports coaching tasks demand precise temporal grounding to identify technique errors and assess form, yet current Vid-LLMs frequently attend to irrelevant frames. We present a dual-pathway self-consistency framework for video-language model training that improves internal frame localization without requiring any frame-level annotations during training and only a small set for offline layer selection. By enforcing attentional agreement between a generation pathway and

an easier verification pathway, we encourage the model to ground its responses in the same relevant video frames regardless of task, using the consistency between these complementary task-views as a self-supervision signal. We show that Vid-LLMs tend to misallocate attention toward irrelevant frames by observing improvements through our self-consistency training across diverse sports understanding benchmarks. We hope this work motivates further exploration of internal attention structure and complementary task-views as a source of self-supervision in video-language models. While our evaluation focuses on sports coaching, a domain where temporal precision and naturally paired tasks make the method particularly well-suited, the underlying principle is domain-agnostic, and extending to other video domains with paired generation-verification structure (e.g., procedural instruction, medical assessment) is a natural direction for future work.

Acknowledgement. This work was supported in part by NSF grant No. 2046853. We thank Sina Malakouti for helpful discussions and feedback on the manuscript.

References

1. Arnab, A., Iscen, A., Caron, M., Fathi, A., Schmid, C.: Temporal chain of thought: Long-video understanding by thinking in frames. arXiv preprint arXiv:2507.02001 (2025)
2. Ashutosh, K., Grauman, K.: Learning skill-attributes for transferable assessment in video. In: Advances in Neural Information Processing Systems (NeurIPS) (2025)
3. Ashutosh, K., Nagarajan, T., Pavlakos, G., Kitani, K., Grauman, K.: Expertaf: Expert actionable feedback from video. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
4. Bao, P., Shao, Z., Yang, W., Ng, B.P., Er, M.H., Kot, A.C.: Omnipotent distillation with llms for weakly-supervised natural language video localization: When divergence meets consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 747–755 (2024)
5. Bao, P., Xia, Y., Yang, W., Ng, B.P., Er, M.H., Kot, A.C.: Local-global multi-modal distillation for weakly-supervised temporal video grounding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 738–746 (2024)
6. Bao, Z., Zhang, L.: Tennistv: Do multimodal large language models understand tennis rallies? arXiv preprint arXiv:2509.15602 (2025)
7. Burgess, J., Wang, X., Zhang, Y., Rau, A., Lozano, A., Dunlap, L., Darrell, T., Yeung-Levy, S.: Video action differencing. In: International Conference on Learning Representations (ICLR 2025) (2025)
8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision (ICCV) (2021)
9. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 397–406 (2021)

10. Chen, H., Huang, H., Yin, X., Shao, D.: Finequest: Adaptive knowledge-assisted sports video understanding via agent-of-thoughts reasoning. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 2909–2918 (2025)
11. Chen, S., Zhu, T., Zhou, R., Zhang, J., Gao, S., Niebles, J.C., Geva, M., He, J., Wu, J., Li, M.: Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. In: Proceedings of the 42nd International Conference on Machine Learning (ICML) (2025)
12. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. pp. 1597–1607. PMLR (2020). https://doi.org/https://doi.org/10.48550/arXiv.2002.05709
13. Cho, J.H., Madotto, A., Mavroudi, E., Afouras, T., Nagarajan, T., Maaz, M., Song, Y., Ma, T., Hu, S., Rasheed, H., Sun, P., Huang, P.Y., Bolya, D., Jain, S., Martin, M., Wang, H., Ravi, N., Jain, S., Stark, T., Moon, S., Damavandi, B., Lee, V., Westbury, A., Khan, S., Krähenbühl, P., Dollár, P., Torresani, L., Grauman, K., Feichtenhofer, C.: Perceptionlm: Open-access data and models for detailed visual understanding. arXiv (2025)
14. Fan, Z., Liu, J., Zhang, Y., Wang, Z., Fung, Y.R., Li, M., Ji, H.: Video-llms with temporal visual screening. arXiv preprint arXiv:2508.21094 (2025)
15. Gao, R., Liu, X., Hu, Z., Xing, B., Xia, B., Yu, Z., Kälviäinen, H.: Fsbench: A figure skating benchmark for advancing artistic sports understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13595–13605 (2025)
16. Google: Gemini 3 flash (2026), https://gemini.google.com/, large Language Model. Accessed: 2026-03-05
17. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., Byrne, E., Chavis, Z., Chen, J., Cheng, F., Chu, F.J., Crane, S., Dasgupta, A., Dong, J., Escobar, M., Forigua, C., Gebreselasie, A., Haresh, S., Huang, J., Islam, M.M., Jain, S., Khirodkar, R., Kukreja, D., Liang, K.J., Liu, J.W., Majumder, S., Mao, Y., Martin, M., Mavroudi, E., Nagarajan, T., Ragusa, F., Ramakrishnan, S.K., Seminara, L., Somayazulu, A., Song, Y., Su, S., Xue, Z., Zhang, E., Zhang, J., Castillo, A., Chen, C., Fu, X., Furuta, R., Gonzalez, C., Gupta, P., Hu, J., Huang, Y., Huang, Y., Khoo, W., Kumar, A., Kuo, R., Lakhavani, S., Liu, M., Luo, M., Luo, Z., Meredith, B., Miller, A., Oguntola, O., Pan, X., Peng, P., Pramanick, S., Ramazanova, M., Ryan, F., Shan, W., Somasundaram, K., Song, C., Southerland, A., Tateno, M., Wang, H., Wang, Y., Yagi, T., Yan, M., Yang, X., Yu, Z., Zha, S., Zhao, C., Zhao, Z., Zhu, Z., Zhuo, J., Arbeláez, P., Bertasius, G., Crandall, D.J., Damen, D., Engel, J., Farinella, G., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C.V., Newcombe, R.A., Park, H.S., Rehg, J.M., Sato, Y., Savva, M., Shi, J., Shou, M.Z., Wray, M.: Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 19383–19400 (2023)
18. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., Valko, M.: Bootstrap your own latent - a new approach to self-supervised learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 21271–21284. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf

19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9726–9735 (2020). https://doi.org/10.1109/CVPR42600.2020.00975
20. He, X., Liu, W., Ma, S., Liu, Q., Ma, C., Wu, J.: Finebadminton: A multi-level dataset for fine-grained badminton video understanding. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 12776–12783 (2025)
21. Hong, F.T., Feng, J.C., Xu, D., Shan, Y., Zheng, W.S.: Cross-modal consensus network for weakly supervised temporal action localization. In: Proceedings of the 29th ACM international conference on multimedia. pp. 1591–1599 (2021)
22. Hou, P., Zhang, J., Huang, D.: Attend and replay: Efficient action understanding in long videos via mechanistic interpretability. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 6602–6611 (October 2025)
23. Kang, S., Kim, J., Kim, J., Hwang, S.J.: See what you are told: Visual attention sink in large multimodal models. In: International Conference on Learning Representations (ICLR 2025) (2025)
24. Kim, J., Kim, K., Seo, S., Park, C.: Compodistill: Attention distillation for compositional reasoning in multimodal LLMs. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=Wa9Bg9b50B
25. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly **2**(1-2), 83–97 (1955). https://doi.org/https://doi.org/10.1002/nav.3800020109, https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109
26. Li, Y., Tang, C., Zhuang, J., Yang, Y., Sun, G., Li, W., Ma, Z., Zhang, C.: Improving llm video understanding with 16 frames per second. In: Proc. ICML. Vancouver (2025)
27. Panchal, S., Bhattacharyya, A., Berger, G., Mercier, A., Böhm, C., Dietrichkeit, F., Pourreza, R., Li, X., Madan, P., Lee, M., et al.: What to say and when to say it: Live fitness coaching as a testbed for situated interaction. Advances in Neural Information Processing Systems **37**, 75853–75882 (2024)
28. Rao, J., Li, Z., Wu, H., Zhang, Y., Wang, Y., Xie, W.: Multi-agent system for comprehensive soccer understanding. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3654–3663 (2025)
29. Rasekh, A., Soula, E.B., Daliran, O., Gottschalk, S., Fayyaz, M.: Enhancing temporal understanding in video-llms through stacked temporal attention in vision encoders. arXiv preprint arXiv:2510.26027 (2025)
30. Su, R., Ouyang, W., Zhou, L., Xu, D.: Improving action localization by progressive cross-stream cooperation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12016–12025 (2019)
31. Team, Q.: Qwen2.5-vl (January 2025), https://qwenlm.github.io/blog/qwen2.5-vl/
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
33. Wang, X., Cheng, F., Wang, Z., Wang, H., Islam, M.M., Torresani, L., Bansal, M., Bertasius, G., Crandall, D.: Timerefine: Temporal grounding with time refining video llm. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5067–5078 (2026)

34. Wang, Y., Li, X., Yan, Z., He, Y., Yu, J., Zeng, X., Wang, C., Ma, C., Huang, H., Gao, J., Dou, M., Chen, K., Wang, W., Qiao, Y., Wang, Y., Wang, L.: Internvideo2.5: Empowering video mllms with long and rich context modeling. arXiv preprint arXiv:2501.12386 (2025)
35. Xia, H., Ge, H., Zou, J., Choi, H.W., Zhang, X., Suradja, D., Rui, B., Tran, E., Jin, W., Ye, Z., Lin, X., Lai, C., Zhang, S., Miao, J., Chen, S., Tracy, R., Ordonez, V., Shen, W., Chen, H.: Sportr: A benchmark for multimodal large language model reasoning in sports. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=cPCGB402ff
36. Xia, H., Yang, Z., Zou, J., Tracy, R., Wang, Y., Lu, C., Lai, C., He, Y., Shao, X., Xie, Z., et al.: Sportu: A comprehensive sports understanding benchmark for multimodal large language models. arXiv preprint arXiv:2410.08474 (2024)
37. Xiao, J., Yao, A., Li, Y., Chua, T.S.: Can i trust your answer? visually grounded video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13204–13214 (2024)
38. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
39. Yang, H., Rao, J., Wu, H., Xie, W.: Soccermaster: A vision foundation model for soccer understanding. arXiv preprint arXiv:2512.11016 (2025)
40. Yi, H., Pan, Y., He, F., Liu, X., Zhang, B., Oguntola, O., Bertasius, G.: Exact: A video-language benchmark for expert action analysis. In: Advances in Neural Information Processing Systems (NeurIPS 2025) (2025)
41. Zhang, J., Khayatkhoei, M., Chhikara, P., Ilievski, F.: Mllms know where to look: Training-free perception of small visual details with multimodal llms. In: International Conference on Learning Representations (ICLR) (2025)
42. Zhang, Q., Singh, C., Liu, L., Liu, X., Yu, B., Gao, J., Zhao, T.: Tell your model where to attend: Post-hoc attention steering for llms. arXiv preprint arXiv:2311.02262 (2023)
43. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (ICLR) (2020), https://openreview.net/forum?id=SkeHuCVFDr
44. Zhang, Y., Qian, S., Peng, B., Liu, S., Jia, J.: Prompt highlighter: Interactive control for multi-modal llms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13215–13224 (2024)
45. Zhang, Z., Yadav, S., Han, F., Shutova, E.: Cross-modal information flow in multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
46. Zou, J., Xia, H., Ye, Z., Zhang, S., Lai, C., Ordonez, V., Shen, W., Chen, H.: Deepsport: A multimodal large language model for comprehensive sports video reasoning via agentic reinforcement learning. arXiv preprint arXiv:2511.12908 (2025)