Understanding Temporal Logic Consistency in Video-Language Models through Cross-Modal Attention Discriminability

Supplementary Material

In this file, we provide additional details as follows:

A. More Details about Analysis
   1. **More Attention Patterns.** This section presents additional attention patterns of key attention heads in TimeChat, further supporting the analysis in Section 2.1 of the main text.
   2. **Correlation Analysis Details.** This section provides the fitting curves and complete distributions of the correlation analysis, as well as further analysis.
   3. **Details of Analysis on EOJ Task.** This section provides additional details on the analysis of the Event Order Judgement (EOJ) task, including the methodology, implementation details, results, and discussion.

B. More Details about Experiments
   1. **Backbone Models Introduction.** This section provides a detailed introduction to the backbone models used in our experiments, along with the reasons for their selection.
   2. **Datasets Introduction and Examples.** This section describes the datasets used in our experiments, including examples to illustrate their content and structure.
   3. **Evaluation Metrics.** This section details the evaluation metrics used to assess the performance.
   4. **Other Baselines and Backbones.** This section discusses additional baselines and backbone models that are considered in our experiments.
   5. **Case Study.** This section presents some case studies to illustrate the effectiveness of our method in enhancing temporal consistency in Video-LLMs.
   6. **Further Ablation Studies on Sensitivity.** This section provides additional ablation studies to further validate the components of our proposed method.

C. More Discussions
   1. **Method Design Related Work.** This section reviews related work on contrastive learning relevant to our method design.
   2. **Limitations and Future Work.** This section discusses the limitations of our current work and outlines potential directions for future research.

A. More Details about Analysis

A.1. More Attention Patterns

In the main text, we presented the attention patterns of key attention heads in TimeChat. To further support the analysis in Section 2.1, we provide additional attention patterns of key attention heads in TimeChat. As shown in Figure 8, we observe that the attention patterns of these heads are similar to those shown in the main text, with a focus on the temporal dimension. This indicates that these heads are indeed crucial for understanding the temporal dynamics of video content.

A.2. Statistical Analysis Details

In the main text, we conduct the correlation analysis of attention discriminability scores and consistency scores in TimeChat. To further support the analysis in Section 2.1, we provide the scatter plot with the least-squares regression line to illustrate the relationship between these two metrics.

As shown in Figure 9, we observe that as the consistency score increases, the distribution of attention discriminability scores shifts to the right. The points clustered on the left side of the x-axis in the figure correspond to difficult samples where the model's predicted IoU is 0.

For rephrased grounding and shifted grounding, the Pearson correlation coefficients between the two metrics are 0.4778 and 0.4788, respectively, with p-values of $\mathbf{5.18 \times 10^{-41}}$ and $\mathbf{3.69 \times 10^{-41}}$, both far below $\mathbf{0.05}$, indicating a statistically significant positive correlation between the two metrics. Since Video-LLMs are inherently complex nonlinear systems, the figure shows that the two metrics are not linearly correlated. However, their positive correlation is statistically significant, as evidenced by the p-values.

A.3. Details of Analysis on EOJ Task

To explore the correlation between attention discriminability and consistency performance on the Event Order Judgement (EOJ) task, we constructed a probing dataset based on the annotations from Charades-STA test subset, containing 1,000 consistent question-answer pairs. Each pair includes a video clip and multiple logically equivalent questions that require the model to determine the order of different events in the video clip. An example of the question-answer pairs is shown in Figure 10.

A.3.1. Detection and Interpretation

Using the same method as in the main text, we first identify the cross-modal attention heads in Qwen2.5-VL that are relevant to the EOJ task. The distribution of these heads is shown in Figure 11. We observe that, consistent with the analysis in the main text, the cross-modal scores in

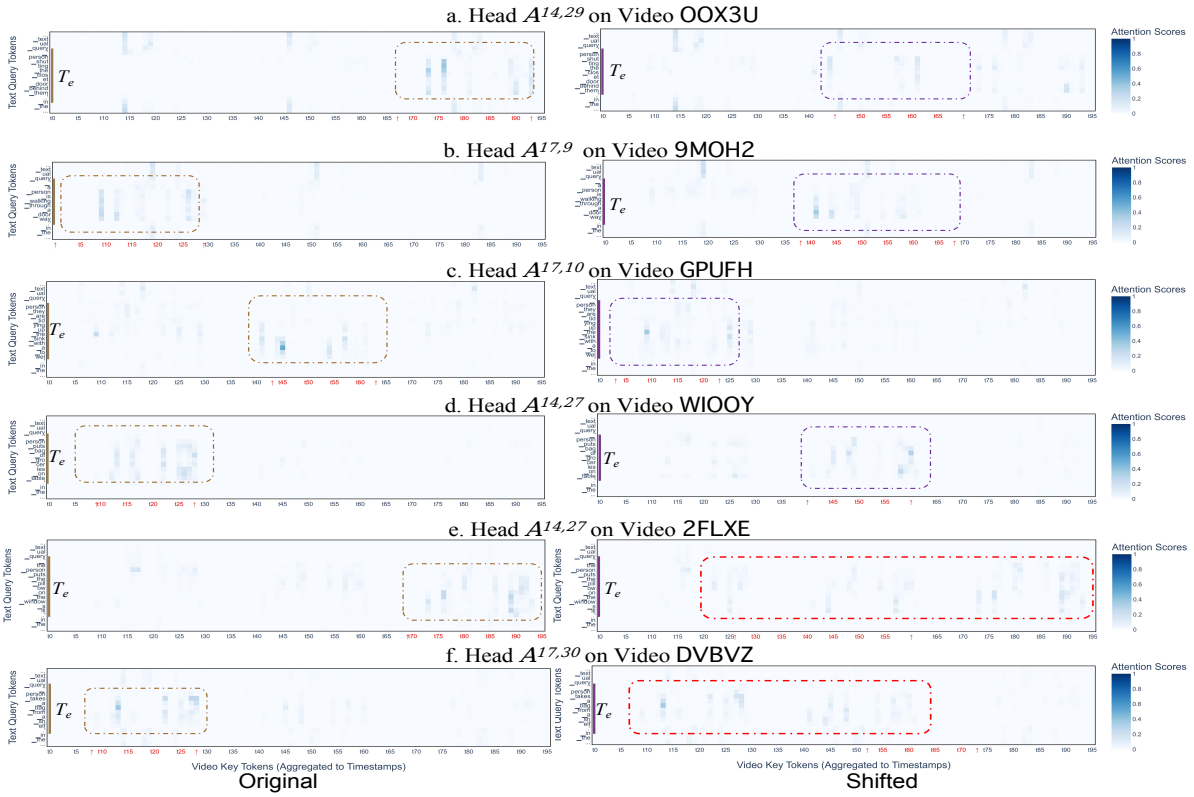

Figure 8. Visualization of attention score distributions for key heads across various samples. Subplots a-d are high discriminability cases, while e and f represent cases where the key heads fail to distinguish correctly.

Qwen2.5-VL are also primarily concentrated in a few heads in the middle layers.

Next, we visualize the attention patterns of key heads on several samples, as shown in Figure 5 in main text. We find that the attention patterns of these heads are consistent with the analysis in the main text, establishing a mapping relationship between the query tokens of event queries and the video key tokens corresponding to the time range of events through attention scores, thus achieving cross-modal temporal relationship alignment.

A.3.2. Correlation Analysis

In the EOJ task, the consistency score c^v is defined as follows:

$$c_v = \frac{\sum_{q \in Q_v} F1(r_{v,q})}{|Q_v|}, \quad v \in \mathcal{D} \tag{5}$$

where c_v indicates the model's **EOJ Consistency Score** for video sample v, $r_{v,q}$ denotes the model's response to question q for video sample v, Q_v refers to the set of questions associated with video sample v, and $\mathcal{D}$ represents the dataset of all video samples.

Considering that the EOJ task involves multiple events, the **EOJ Attention Discriminability Score** is defined as the Kullback-Leibler divergence [11] between the average attention score distributions of two events as follows:

$$P_e^{h,v} = \frac{\sum_{q \in T_e} A_{q,V}^{h,v}}{\sum_{k \in V} \sum_{q \in T_e} A_{q,k}^{h,v}}, \tag{6}$$

$$S_{eoj,disc.}^{h,v} = P_{e_1}^{h,v} \log \frac{P_{e_1}^{h,v}}{P_{e_2}^{h,v}} + P_{e_2}^{h,v} \log \frac{P_{e_2}^{h,v}}{P_{e_1}^{h,v}}, \tag{7}$$

$$S_{eoj,disc.}^{v} = \frac{1}{|H_t|} \sum_{h \in H_t} S_{eoj,disc.}^{h,v}, \tag{8}$$

where $P_e^{h,v}$ denotes the average attention score distribution of all text tokens of event e to the set of all visual tokens V, T_e represents the set of text tokens of event e, $A_{q,V}^{h,v}$ is the attention score from text token as query token q to all visual tokens V as key tokens, $S_{eoj,disc.}^{h,v}$ is the EOJ temporal discriminability score of attention head h for video sample v containing two events e_1 and e_2, H_t means the set of the top t cross-modal score attention heads, and $S_{eoj,disc.}^{v}$ is the average eoj temporal discriminability score of the top t cross-modal attention heads for video sample v.

The correlation analysis of attention discriminability scores and consistency scores is shown in Figure 13 as a scatter plot with the least-squares regression line. We observe that as the consistency score increases, the attention

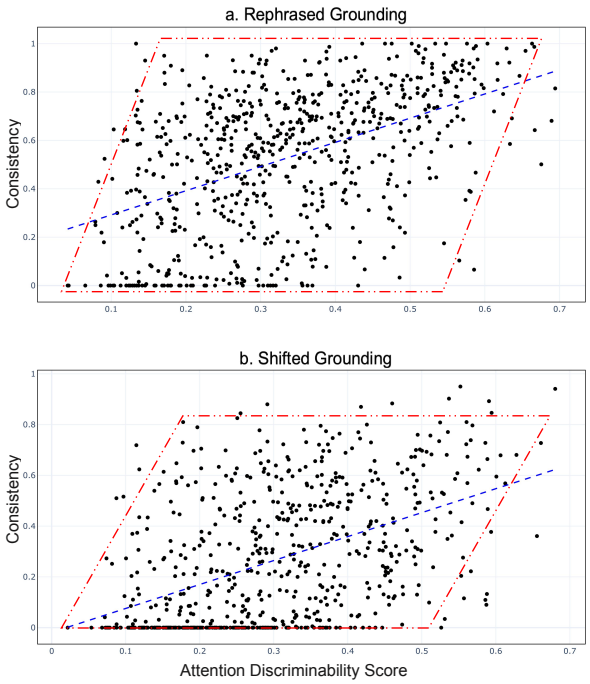

Figure 9. Scatter plot showing the relationship between the consistency score and the attention discriminability score, with a least-squares regression line included.

Models	Consistency Metrics				
	G	R-Ground	S-Ground	H-Verify	C-Verify
Charades-CON					
GPT-4o	28.5	21.2 (74.3)	9.3 (32.8)	17.8 (62.4)	20.3 (71.3)
gemini-1.5	34.6	29.7 (85.7)	24.8 (71.7)	22.8 (65.8)	24.5 (70.8)
ActivityNet-CON					
GPT-4o	26.8	18.1 (67.5)	10.4 (38.8)	16.5 (61.7)	18.4 (68.8)
gemini-1.5	37.8	30.8 (81.4)	24.8 (65.6)	22.4 (59.3)	26.8 (70.8)

Table 7. Results of other models without tuning, listed for reference only [8].

discriminability score also increases significantly, indicating a clear positive correlation between attention discriminability and consistency performance. The corresponding violin plot is shown in Figure 12.

In conclusion, our analysis is validated across both the Qwen2.5-VL model and the EOJ task, underscoring the robustness and generalizability of our conclusions. This provides compelling evidence that attention temporal discriminability is a key factor of performance for Video-LLMs in general temporal understanding tasks.

EOJ Probing Dataset

EARLIER THAN
Visual Input: The frames of 4H64T.mp4
Question: Does event 'person opens a refrigerator' occur earlier than event 'person eating some leftovers from a take-out carton' in the video?
Answer: Yes.
Question: Does event 'person eating some leftovers from a take-out carton' occur earlier than event 'person opens a refrigerator' in the video?
Answer: No.

LATER THAN
Visual Input: The frames of 4H64T.mp4
Question: Does event 'person eating some leftovers from a take-out carton' occur later than event 'person opens a refrigerator' in the video?
Answer: Yes.
Question: Does event 'person opens a refrigerator' occur later than event 'person eating some leftovers from a take-out carton' in the video?
Answer: No.

AFTER
Visual Input: The frames of 4H64T.mp4
Question: In the video, after event 'person opens a refrigerator', does event 'person eating some leftovers from a take-out carton' occur?
Answer: Yes.
Question: In the video, after event 'person eating some leftovers from a take-out carton', does event 'person opens a refrigerator' occur?
Answer: No.

BEFORE
Visual Input: The frames of 4H64T.mp4
Question: In the video, before event 'person eating some leftovers from a take-out carton', does event 'person opens a refrigerator' occur?
Answer: Yes.
Question: In the video, before event 'person opens a refrigerator', does event 'person eating some leftovers from a take-out carton' occur?
Answer: No.

Figure 10. An example of the EOJ Probing dataset. Each sample contains two events with a clear order, and four sets of eight logically equivalent questions are constructed for their order relationship.

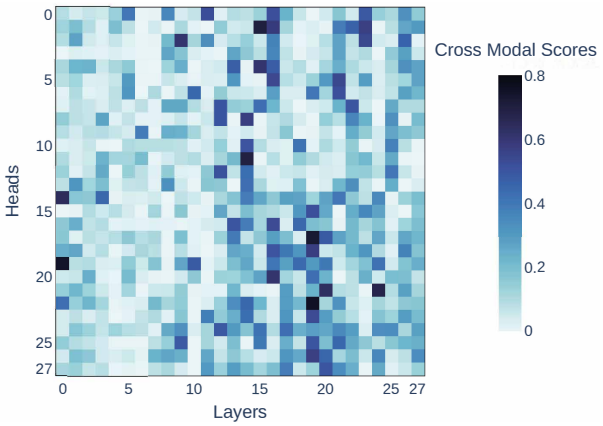

Figure 11. The distribution of cross-modal attention heads in Qwen2.5-VL. The x-axis represents the attention layer index, and the y-axis represents the head index.

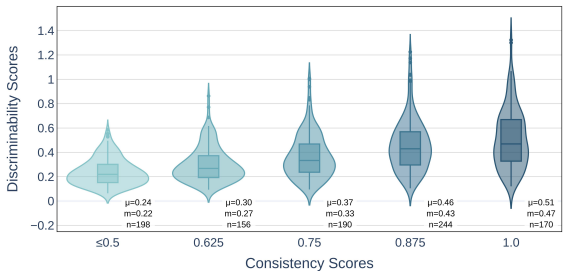

Figure 12. Violin plot of analysis results on EOJ task.

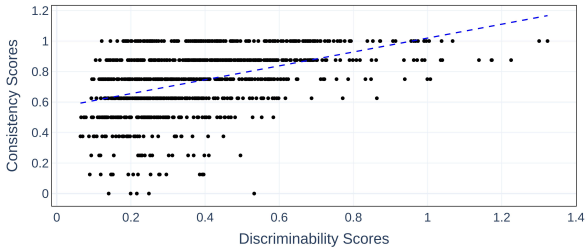

Figure 13. Scatter plot showing the relationship between the consistency score and the attention discriminability score on EOJ task, with a least-squares regression line included.

B. More Details about Experiments

B.1. Backbone Models Introduction

Qwen2.5-VL [1] is a representative Video-LLMs featuring a standard cross-modal architecture. We select Qwen2.5-VL to further validate our findings from the previous section. **Video-LLaMA** [28] is a classic video-LLM using Qformer and video Qformer for video-text alignment. **TimeChat** [19], a SOTA video-LLM for the VTG task, integrates timestamp information into its architecture, enhancing its ability to understand the temporal dynamics of video content. We include Video-LLaMA and TimeChat to ensure a fair comparison with the current SOTA method.

B.2. Datasets Introduction and Examples

TimeIT *TimeIT* is a dataset proposed by Ren et al. [19] for the various temporal understanding tasks including description, video grounding and temporal reasoning. The splits Charades-VTG and ActivityNet-VTG of it is used in our experiments. These two splits contain 9,848 video clips with 12,408 question-answer pairs and 14,420 video clips with 35,692 question-answer pairs respectively.

VTune *VTune* is a dataset reported as a method in the original paper [8]. By constructing event verification queries and temporal verification queries annotation data, it effectively enhances the consistency of Video-LLMs from a view of synthetic data augmentation. Two splits Charades-VTune and ActivityNet-VTune contains 9,848 video clips with 99,244 question-answer pairs and 14,420 video clips with 205,510 question-answer pairs.

TimeIT
Visual Input: The frames of WIOOY.mp4
Question: Localize the visual content described by the given textual query 'A person starts cooking' in the video, and output the start and end timestamps.
Answer: The given query happens in 12.4 to 27.0 seconds.

VTune
(Event Verification Query)
Visual Input: The frames of WIOOY.mp4
Question: Does the event 'The person is eating dinner at a table' not happen from 12.4 to 27.0 seconds in the video?
Answer: Yes, the event 'The person is eating dinner at a table' does not happen from 12.4 to 27.0 seconds in the video. We cannot see a person eating dinner; the individual is cooking.

(Temporal Verification Query)
Visual Input: The frames of WIOOY.mp4
Question: Is the event 'A person starts cooking' present from 0 to 10 seconds in the video?
Answer: No, we can see the event 'A person starts cooking' from 12.4 to 27.0 seconds.

Figure 14. Examples from the training datasets (TimeIT and VTune) used in our training experiments.

CON *Charades-CON* and *ActivityNet-CON* are two benchmarks proposed by Jung et al. [8] to evaluate the consistency of Video-LLMs in temporal understanding tasks. These benchmarks contain 500 video clips with 707 question-answer pairs and 1,422 question-answer pairs.

The examples of TimeIT and VTune are shown in Figure 14. These two datasets are used for training, while the last two datasets are used for evaluation in our experiments.

B.3. Evaluation Metrics

B.3.1. Evaluation Metrics Details

Following the previous works [8], we use these metrics to evaluate the performance of our method:

- **IoU**: The ratio of the intersection over the union between the predicted time range and the ground-truth time range.
- **Accuracy**: The percentage of questions that can be answered correctly by the model. We use it evaluate absolute performance on the verification tasks (H-V. and C-V.) of CON datasets.
- **R@1,IoU=0.5**: The percentage of the IoU of prediction is greater than 0.5. We use it to evaluate the absolute per-

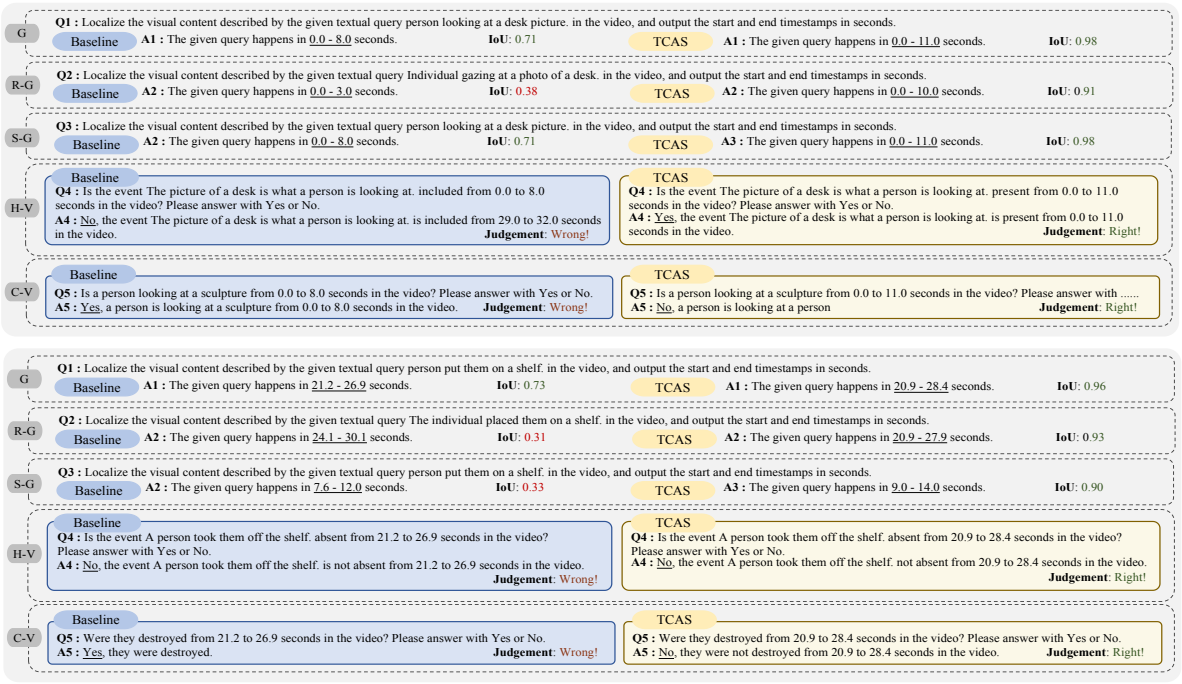

Figure 15. Responses of TimeChat for sample 9T11N (Top) and T0HLX (Bottom) before and after TCAS enhancement.

formance of the model on all grounding tasks.

- **R@1,IoU=0.7**: The percentage of the IoU of prediction is greater than 0.7. We use it to evaluate the absolute performance of the model on all grounding tasks.
- **mIoU**: The mean IoU of all questions in the dataset.
- **Consistency**: All Consistency metrics are defined as the original grounding R@1,IoU=0.5 times corresponding absolute performance metrics.

B.3.2. Metrics Rationality

For analysis in Section 3, we use a sample-wise consistency score (IoU product) for instance-level correlation. In our main experiments (Sec. 5), we follow Jung et al.'s protocol [8], reporting R@0.5, Acc, and their product with the original grounding score as Appendix Section B.3.1, rather than sample-wise score for strictly fair comparison with baselines. For intervention, we report per-setting IoU (Ori/R./S.) to isolate subtask effects.

The result using sample-wise consistency score (the analysis metric) is shown in Table 8. A simple intervention on a small number of attentional heads can slightly improve the consistency score, while low-intensity random intervention leads to a rapid decline in performance. Gains may look limited because most components are frozen and cannot co-adapt to the intervention, constraining immediate gains. TCAS mitigates this by retraining, allowing the improved attention discriminability to better translate into performance, further confirming the effectiveness of our method TCAS in enhancing temporal consistency.

Table 8. Results of sample-wise consistency score defined as the product of IoU and original grounding score. The α means the intervention intensity. "Random Int." and "Ground Truth Int." mean random intervention and intervention based ground-truth.

	SFT w/o Int.	**Random Int.**			**Ground Truth Int.**					**TCAS**
α		0.05	0.1	0.15	0.2	0.4	0.6	0.8	1	-
$c_{rg.}$	65.4	39.6	2.1	0	66.0	65.4	64.1	62.9	61.2	**70.0**
$c_{sg.}$	33.1	17.2	1.8	0	33.5	33.7	34.0	34.5	34.7	**37.5**

B.4. Results of Other Baselines and Backbones

We also compare our method with the following closed-source models: GPT-4o [16] Gemini-1.5 Flash [21] in table 7. It is worth noting that these results are reported in [8] and are listed here for reference only.

Furthermore, we conducted comparative experiments on the general VTG task using Qwen2.5-VL and Video-LLaMA, with the results presented in Table 9. In contrast to the improvements seen with TimeChat, our method does not consistently enhance performance on the general VTG task for these models. However, the general localization performance remains stable or slightly improves with the enhanced consistency, which we consider acceptable, as the primary motivation of our work is to improve consistency in an interpretable manner.

Table 9. More comparison results for VTG performance on Charades-STA and ActivityNet-Caption.

Method	Charades-STA		ActivityNet-Cap.	
	R@1,0.5	R@1,0.7	R@1,0.5	R@1,0.7
videollama (w. SFT)	37.1	20.1	34.3	19.1
videollama (w. TCAS)	37.5	20.1	35.2	20.0
Qwen2.5vl (w. SFT)	27.3	14.1	16.5	7.8
Qwen2.5vl (w. TCAS)	32.4	16.6	16.5	7.9

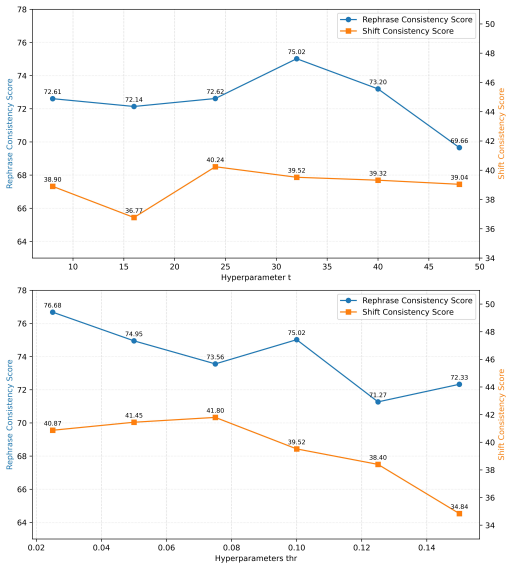

Figure 16. Performance sensitivity analysis of hyperparameters t (top subplot) and thr (bottom subplot).

B.5. Case Study

In this section, we present more prediction examples to illustrate the effectiveness of our method in enhancing temporal consistency in Video-LLMs. Our model successfully grounds key events and answers questions about the video, demonstrating improved temporal capabilities and consistency compared to baseline models.

B.6. Further Ablation Studies on Sensitivity

We conducted a more fine-grained sensitivity analysis for the relatively sensitive hyperparameters t and thr. The performance variation with respect to these hyperparameters is shown in Figure 16. From the figure, we observe that: 1) As t increases, both consistency scores exhibit an overall trend of first increasing and then decreasing, emphasizing the need for precise regulation of attention heads, as excessive intervention on heads weakly related to temporal discriminability can have negative effects. 2) When thr is small, both Rephrase Consistency Score and Shift Consistency Score are relatively high; however, when thr exceeds 0.1, performance rapidly declines. Because larger values of thr result in too few tokens being adjusted, which fails to effectively enhance attention discriminability.