

Video-CoE: Reinforcing Video Event Prediction via Chain of Events

Supplementary Material

A. Additional Results

We additionally include comparisons with RL-based methods and traditional state-of-the-art approaches.

Table 1. Additional results, *denotes traditional method.

Model	FutureBench \uparrow		AVEP \uparrow	
	AVG	Verb	Noun-F1	Action-F1
VideoChat-R1	46.59	8.31	42.22	3.95
Video-R1	67.47	9.47	47.04	4.32
Ours	75.00	18.75	64.03	9.88
EventFormer*	-	22.71	46.24	7.69

B. Attention Score Calculation Process

To quantitatively evaluate the MLLMs’ attention to visual information during the video event prediction task, we compute the attention weights assigned to both visual and text tokens. We record the attention score matrices of every head at every layer during inference on the test set. Since existing open-source MLLMs contain a large number of layers and heads, and we did not observe notable differences across them in our experiments, we visualize the average attention scores aggregated over all LLM layers and heads. To control for the variation in the number of tokens generated by different models, we visualize the attention score distribution from option tokens to all other tokens. This provides a direct insight into how each model allocates attention between visual and textual information during event prediction, while also enabling an unbiased comparison of their visual attention.

C. Similarity Reward Computation

For an event chain of length n , where the video segments are represented as $[clip_1, clip_2, \dots, clip_n]$ and the corresponding descriptions as $[\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_n]$, we use the similarity model $f_\theta(\cdot)$ to embed the video feature v and text feature t of the event chain:

$$v_j = f_{\theta_{visual}}(clip_j), t_j = f_{\theta_{text}}(\mathcal{D}_j). \quad (1)$$

Thus, the similarity reward r_s is obtained by averaging the similarities between the video and text features of the events:

$$r_s = \frac{1}{n} \sum_{j=1}^n (\text{sim}(v_j, t_j)), \quad (2)$$

where $\text{sim}(\cdot)$ denotes the similarity computation function, which typically refers to cosine similarity.

In the experiments described in the main paper, we explore two different approaches for calculating the similarity reward. The first method directly uses a video-text alignment model to compute the video features of cropped video event segments and the text features of their corresponding descriptions. For this approach, we follow the official recommendation and use a frame rate of 8 frames for sampling. The second method employs an image-text alignment model, where we extract image frames from the video segments, calculate the similarity between each image feature and the text feature, and then average these similarity scores to obtain the overall similarity between the video event segment and its textual description. We use a frame rate of 8 frames for sampling as well.

D. Video Event Prediction Benchmarks

Futurebench. Futurebench is a benchmark specifically designed to evaluate the video event prediction capability of MLLMs, featuring both SFT and GRPO training datasets. This benchmark collects video data from various perspectives, different lengths, and types, offering a comprehensive assessment of event prediction performance across diverse scenarios. To evaluate the event prediction ability of MLLMs from multiple angles, Futurebench categorizes the event prediction tasks into four types:

- **1-Hop:** The model predicts a single future event that directly links the observed scenes to the final one, corresponding to a standard Next Event Prediction (NEP).
- **2-Hop:** The model infers a sequence of two consecutive future events, requiring a short chain reasoning process that sequentially connects the observed scenes to the final event.
- **3-Hop:** The model predicts three consecutive future events, significantly increasing task complexity by necessitating deeper causal reasoning across a longer temporal span.
- **Interp.:** The model must infer multiple non-consecutive future events, given a set of partially observed scenes that include intermediate anchor events.

In this benchmark, we construct **2,000** CoE-SFT samples for CoE-SFT training and additionally utilize the provided **2,000** reinforcement learning samples as training data for CoE-GRPO. As the training set contains only 1-Hop and 2-Hop event prediction samples, performance gains on the 3-Hop and Interp. metrics provide a reliable measure of the model’s generalization capability on VEP task.

AVEP. Action-centric Video Event Prediction (AVEP) is a benchmark specifically designed for evaluating the video

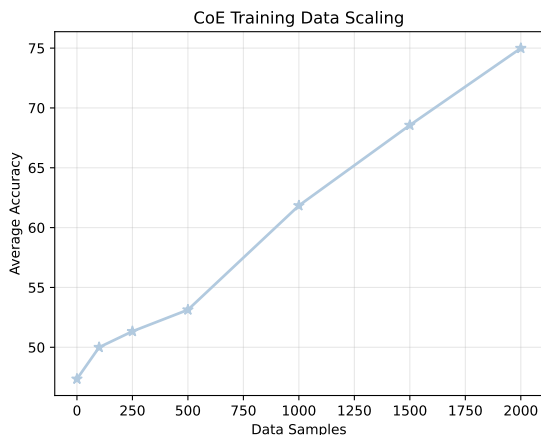


Figure 1. The data scaling curve of CoE

event prediction capabilities of models. This benchmark primarily focuses on the events themselves, providing a comprehensive and fine-grained assessment of the model’s performance on video event prediction. Video events are decomposed into event arguments, and the model’s ability to predict future events is evaluated at the argument level. The evaluation of a model’s event prediction performance on AVEP is based on the following key aspects:

- **Verb Accuracy:** This metric measures the accuracy of the model in predicting the trigger verb of future events. As verbs are the core triggers of events, they are crucial components in event construction, and their prediction reflects the model’s logical reasoning ability.
- **Noun Metrics:** This set of metrics assesses the model’s ability to predict the participants in future events, including the subject, object, and tool. These metrics reflect the model’s consistency in role prediction, evaluating the logical coherence of the event arguments.
- **Action Metrics:** These metrics evaluate the noun prediction performance when the verb is correctly predicted, providing a direct indication of the model’s video event prediction ability.

The AVEP dataset provides the ground truth but does not include the SFT data required for training MLLMs. Therefore, we use the **5,000** constructed CoE-SFT samples for supervised fine-tuning and experiments. In this benchmark, we select **5,000** samples from the provided dataset as the training data for GRPO and CoE-GRPO, and evaluate the model on the entire validation and test sets of the benchmark.

E. Training Data Scaling of CoE

Using the benchmark-provided training data, we train the model to follow the CoE paradigm, and observe strong im-

provements on VEP task, highlighting the data efficiency of our method.

To further investigate the effect of data scale, we conduct a data-scaling study. As shown in the Fig. 1, directly applying the CoE paradigm without any CoE-specific training leads to a performance drop, indicating that the model is unable to effectively leverage visual information to construct the logical connections to future events. When a small amount of training data is provided, performance begins to improve steadily, suggesting that the model is gradually acquiring the CoE reasoning pattern. As the data size increases, performance improves rapidly, demonstrating the strong data efficiency of our approach. Notably, although we train the model on only 2,000 samples from FutureBench, the upward trend in the scaling curve remains far from saturated. This indicates that additional data would likely yield further gains, highlighting the strong scalability and continued potential of our proposed method .

F. Details of Judge Model Evaluation

To more accurately assess MLLMs’ event prediction capabilities in real-world applications, the judge model evaluation is designed to reflect an open-set prediction setting. Specifically, we remove all answer options and require the model to directly reason about the observed video and predict the most plausible future event. In the evaluation, we provide the judge model with the video, question, and reference answer. The judge model then evaluates each output from two perspectives: (i) the logical consistency and soundness of the reasoning, and (ii) the correctness of the predicted event. Directly scoring model outputs may lead to inconsistencies or hallucination from the judge model, undermining evaluation fairness. To address this, we employ a group-wise comparison protocol and report the win rate as the evaluation metric, which yields a more reliable assessment.

We provide a comprehensive example of the judge model evaluation, as shown in Fig. 2. In this evaluation, it is observed that the models trained with the CoE-SFT method produce reasoning processes that are more closely aligned with the video content, and their reasoning is visually grounded, clear, and concise. In contrast, other methods either fail to focus on the visual content during reasoning or provide predictions that lack logical consistency. In other test samples, we find that both CoE-SFT and CoE-GRPO methods are able to deliver reasonable reasoning processes and accurate predictions in the absence of options.

G. Examples of CoE

As illustrated in Figs. 3 to 8, we present several randomly selected prediction examples generated by our proposed method. From the constructed event chains in these results

(marked as blue in the figures), it can be observed that our approach enables the model to perform fine-grained temporal modeling of the input videos and efficiently improves the utilization of visual information. Notably, as the overall video duration varies, the granularity of event segmentation adaptively adjusts, while the length of the generated event chains remains relatively stable. In addition, the textual descriptions produced for each event are generally consistent with the corresponding video segments.

The followed reasoning process (marked as green in the figures) exhibited in these examples demonstrates that the model can logically infer future events based on the details present in the video. This not only shows that our proposed method enables the model to effectively establish logical connections between the video and future events, but also indicates that after CoE-GRPO training, the model retains the logical reasoning capabilities learned during CoE-SFT.

However, as illustrated in the Figs. 9 and 10, we also observed some bad cases, though their occurrence is extremely rare—approximately three instances out of one thousand samples. In these cases, the model fails to generate timestamps correctly according to the given instructions. Nevertheless, this issue has minimal impact on the model’s event descriptions and prediction results.

H. Prompt Templates

Fig. 12 illustrates the prompt template we use for training and inference of all MLLMs. We also present the prompt used for CoE-SFT data generation and judge model evaluation in Fig. 13

I. Visualizations of Attention Increase in Visual Tokens

Here are some visualizations of the attention differences between the post-trained models and the vanilla model (Qwen2.5-VL-7B-Instruct) for visual tokens, as shown in Fig. 11. The portions of the curves above 0 indicate an increase in attention. It is clear that both the CoE-GRPO and CoE-SFT methods effectively enhance the model’s focus on visual information. However, while both vanilla GRPO and vanilla SFT improve the model’s event prediction performance, they fail to adequately address the issue of insufficient visual information utilization, thus limiting the overall efficiency of performance improvement. Additionally, when we directly prompt the model to predict future events in a CoT manner without training, it does not significantly improve the model’s attention to visual content.

J. Limitations and Future Works

We hope that this work can serve as a foundation and provide inspiration for future explorations of MLLMs’ capabilities in video event prediction. Below, we also summarize

the limitations and outline several potential directions for future research:

- **Temporal Localization Ability.** Our proposed method imposes certain requirements on the model’s temporal localization capability. Although most current MLLMs demonstrate strong temporal localization performance, there is still considerable room for improvement in accuracy. We believe that further enhancement of this capability could lead to additional performance gains for our method.
- **Structures of Historical Events.** We explored the use of event-chain construction to strengthen the model’s fine-grained temporal modeling of videos. However, we only investigated relatively basic forms of event chains. More complex formulations for modeling historical event structures—such as relation-aware event chains or event graphs—remain promising directions for future research.
- **Performance on Other MLLMs and Tasks.** Since our approach does not impose specific requirements on the base model, we believe that the CoE paradigm should be applicable to all MLLMs. However, due to resource constraints, we were unable to conduct experiments across all available models. Nevertheless, the exceptional performance of our method on VEP demonstrates its potential for application to other tasks as well. Exploring the applicability of CoE to other video-related tasks presents an interesting avenue for future research.



Question:

“Based on the given video, predict the sequence of future events and fill in the missing event slot with the next major scene that would lead into the final thank you and engagement prompts: 1. [?] 2. A red-curtain sequence displaying messages such as ‘Please Subscribe!’, ‘Thumbs Up’, and ‘Leave a Comment’. Which of the following potential events best fits the gap represented by [?]?” *Options-*

CoE-GRPO:

<event>Time:....., Des:.....</event>\n<event>Time:....., Des:.....</event>\n<event>Time:....., Des:.....</event>

<think>Given the sequence of events, it is clear that the video is structured to build engagement with the audience. The initial scenes set up the context and introduce the main subject, followed by a series of frames encouraging viewer interaction. The next logical step in this sequence would be to transition from the interactive content to a closing segment where the creator thanks the viewers and encourages them to take action, such as subscribing or leaving comments.</think>

The potential event that best fits the gap represented by [?] is a scene where the person concludes their presentation or demonstration, thanking the viewers for watching and encouraging them to subscribe, like, and comment. This would naturally lead into the red-curtain sequence displaying messages such as ‘Please Subscribe!’, ‘Thumbs Up’, and ‘Leave a Comment’.

GRPO:

<think>\n\nThe video appears to be structured around a typical YouTube video format where the creator engages with their audience through various scenes before concluding with a call-to-action. Given the context provided, the missing event slot [?] should logically follow the initial setup or introduction of the content but precede the red-curtain sequence that encourages viewer interaction.\n\nA common structure for YouTube videos includes an opening scene that sets the tone or introduces the topic, followed by a main content segment, and then a closing segment that includes calls-to-action. Since the red-curtain sequence is designed to engage viewers after the main content, the missing event slot [?] should likely be a scene that transitions from the main content to the closing segment.\n\nConsidering this, the most fitting event for the gap would be a scene that summarizes the key points discussed in the main content or provides a brief recap. This would help prepare the audience for the closing segment and ensure they understand the value of the content before being prompted to subscribe, like, or comment.\n\n</think>

The missing event slot [?] should be filled with a summary or recap of the main content.

CoE-SFT:

<think> The video transitions from a hookah demonstration to a driving scene at night, suggesting a shift in location or context. This could indicate a new segment or a transition to a different setting before concluding with the final thank you and engagement prompts. Therefore, the next major scene likely involves a return to the hookah theme or a related activity, fitting logically before the closing messages. </think>

A person preparing hookah tobacco and charcoal, followed by a scene showing the preparation of the hookah itself.

SFT:

<think> Event 4 (the red curtain sequence) is already provided in the final scene, so the missing event must occur between the observed scenes and this final segment. Given the narrative progression from the initial setup to the final call-to-action, the most logical intermediate event is the introduction of the product or service being promoted. This aligns with the observed scenes where the man is seen in various settings, including a kitchen and a living room, which suggests he is demonstrating or discussing something related to the product. The red curtain sequence then serves as a transition to the final call-to-action. </think>

The missing event is the demonstration or discussion of the product/service, which logically bridges the observed scenes and the final red curtain sequence.

Instruct:

<think>\n\nThe video starts with a person in a kitchen setting, possibly preparing food or cooking. This is followed by a transition to a black screen with text, which seems to be an introduction or title card. After this, there’s a scene with a person in a different setting, possibly a studio or a room with a desk, where they appear to be speaking or presenting something. The sequence ends with a red-curtain sequence displaying messages encouraging viewers to subscribe, like, and comment.\n\nGiven the structure of the video, it’s logical to assume that the missing event slot [?] should follow the initial kitchen scene but precede the studio or presentation scene. This could be another segment where the person continues their activity, perhaps explaining a recipe or demonstrating a cooking technique. This would provide a smooth transition from the kitchen to the more formal presentation style seen later.\n\n</think>

The missing event slot [?] should be a continuation of the kitchen scene, possibly showing the person preparing or demonstrating a specific step in the cooking process. This would logically lead into the subsequent scenes where the person moves to a more formal setting for a presentation or explanation.\n

Judge Model :

CoE-GRPO assumes the video follows a typical YouTube format with an introduction, main content, and a conclusion. It suggests the missing event is a summary or recap of the main content. The assumption about the video’s structure is reasonable, **but it doesn’t consider the specific content shown in the video, such as the hookah demonstration and driving scene.**

GRPO also assumes a typical YouTube video structure and suggests the missing event is a summary or recap of the main content. Similar to EGRPO, the reasoning is plausible **but lacks specificity regarding the actual content of the video.**

CoE-SFT considers the transition from the hookah demonstration to the driving scene and suggests the missing event involves further preparation of the hookah. **This reasoning is more aligned with the actual content of the video, as it directly addresses the hookah-related activities shown.**

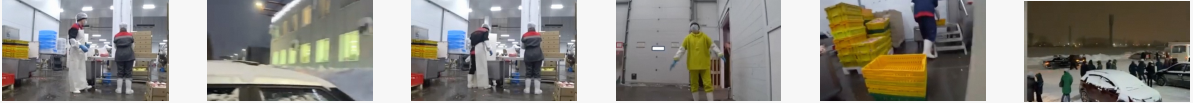
SFT assumes the video is promoting a product or service and suggests the missing event is a demonstration or discussion of the product. **While this reasoning is somewhat relevant, it does not specifically address the hookah-related activities depicted in the video.**

Instruct assumes the video is about cooking and suggests the missing event is a continuation of the kitchen scene. **This reasoning is not accurate, as the video clearly focuses on hookah-related activities rather than cooking.**

Best Model: CoE-SFT. Reason: **CoE-SFT provides the most accurate and relevant response by considering the specific content of the video, particularly the hookah demonstration and the transition to the driving scene.** It correctly identifies that the missing event should involve further preparation of the hookah, which logically fits the sequence of events shown in the video."

Figure 2. An example of judge model evaluation.

Input Video



Question:
 "Based on the given video, predict future events and fill in the blanks in the following sequence: 1. A group of people gathers around snow-covered cars in a parking lot under streetlights. 2. [?] 3. An exterior view of a factory with a Russian sign indicating deep meat processing and production is shown with snow surrounding it. 4. [?] 5. A close-up shot of food being served with the caption 'Food at the factory is free' is displayed. Which option best completes the missing events with the correct intermediate scenes that occur after the observed scenes?"

Options:


- A. 2. An interior view showing a car's dashboard and steering wheel as it drives on a snowy road, 4. A scene inside a locker room where a person adjusts their clothing in preparation for work.
- B. 2. A scene inside a locker room where a person adjusts their clothing in preparation for work, 4. An interior view showing a car's dashboard and steering wheel as it drives on a snowy road.
- C. 2. A display of a white truck with 'АГРОХОЛДИНГ' in red letters parked in a snowy area, 4. A scene inside a locker room where a person adjusts their clothing in preparation for work.
- D. 2. An interior view showing a car's dashboard and steering wheel as it drives on a snowy road, 4. A display of a white truck with 'АГРОХОЛДИНГ' in red letters parked in a snowy area.

CoE:

<event>Time:00:00-00:05, Des: A man in a red shirt speaks directly to the camera.</event>
 <event>Time:00:05-00:10, Des: Two individuals in protective gear are shown working in a factory environment.</event>
 <event>Time:00:10-00:15, Des: Text appears on screen asking if working for 30,000 rubles per month at a meat processing plant is worth it.</event>
 <event>Time:00:15-00:20, Des: The same two individuals continue working in the factory, with one person moving boxes and another handling machinery.</event>
 <event>Time:00:20-00:25, Des: A man in a black jacket is shown speaking while holding a phone, discussing the cost of a different apartment.</event>
 <event>Time:00:25-00:30, Des: An exterior view shows a group of people gathered around snow-covered cars in a parking lot under streetlights.</event>
 <think>The video transitions from a discussion about working conditions in a factory to a scene of people gathering around snow-covered cars, suggesting a shift in location or context. The next logical step would be to show an interior view of the factory or related activities, followed by a scene that might involve the workers preparing for their day, which aligns with the typical workflow in such environments. The truck mentioned in the options could be related to transportation or logistics within the factory setting.</think>
 <answer>A</answer>

Figure 3. An example of CoE paradigm.

Input Video



Question:
 "Based on the given video, predict future events that would occur after the observed scenes (the filling, heating, and stirring of the substance behind the counter) and before the final scene. Complete the following sequence with the correct intermediate events: 1. The video shifts to the person displaying an orange, flexible flat sheet to the camera to emphasize its texture. 2. [Missing Event A] 3. The scene then focuses on a close-up of several skewers holding neatly arranged orange, rectangular food items. 4. [Missing Event B] 5. Finally, a detailed close-up shows a dark, glossy circular object being manipulated over a white bowl with a wooden skewer. Which pair of events best fits as Missing Event A and Missing Event B, respectively, based on the full video narrative?"

Options:

- A. The person interacts with a customer by handing over one of the orange sheets, then later explains or demonstrates product details to a colleague.
- B. The person quickly glances at the packaged snacks before moving on, then rearranges the skewers without further explanation.
- C. The person examines the flat sheet without engaging in any customer interaction, then briefly picks up a bag of small packaged snacks.
- D. The person ignores the flat sheet and starts organizing the counter, then later fixes the position of the skewers without any interaction.

CoE:

<event>Time:00:00-00:05, Des: A white powdery substance is poured into a metal container.</event>
 <event>Time:00:05-00:10, Des: A person in a striped shirt is seen stirring a pot on a stove.</event>
 <event>Time:00:10-00:15, Des: The person pours a dark liquid from a pot into a white bowl using a wooden skewer.</event>
 <event>Time:00:15-00:20, Des: The person places the bowl with the dark liquid on the counter next to various packaged snacks.</event>
 <event>Time:00:20-00:25, Des: The person picks up an orange, flexible flat sheet and displays it to the camera.</event>
 <event>Time:00:25-00:30, Des: The person arranges several skewers with neatly arranged orange, rectangular food items on the counter.</event>
 <event>Time:00:30-00:35, Des: The person manipulates a dark, glossy circular object over a white bowl using a wooden skewer.</event>
 <think>The video sequence shows a series of actions related to preparing and presenting food items. The first part involves pouring and stirring substances, which suggests a cooking or food preparation process. The middle part shows the person handling an orange flat sheet and arranging skewers with food items, indicating a display or presentation phase. The final part involves manipulating a dark, glossy object, likely as part of the food preparation or serving process. Given this context, the missing events should logically follow the sequence of actions and interactions. Option A provides a logical continuation where the person interacts with the flat sheet and then explains or demonstrates product details to a colleague, fitting the narrative flow.</think>
 <answer>A</answer>

Figure 4. An example of CoE paradigm.

Input Video

Question:
 “Based on the given video, predict future events and fill in the missing steps in the following sequence: 1. A Pyrex cup is shown being filled with red wine vinegar, water, and Italian dressing mix. 2. [?] 3. The creamy dressing is poured over the bowl of pasta salad and mixed thoroughly. 4. [?] 5. A block of feta cheese is crumbled and added to the pasta salad as the final step. Which of the following options correctly predicts the missing events in the proper order?”

Options:

- A. A hand stirs in vegetable oil to blend the dressing ingredients; A container of Mediterranean spice mix is sprinkled over the salad.
- B. A hand stirs in mayonnaise to create a smooth dressing; A hand adds extra olives to the salad.
- C. A hand pours vegetable oil over the dressing, forming a distinct layer and then mixes it thoroughly; A hand sprinkles Mediterranean spice mix before finishing with crumbled feta cheese.
- D. A hand mixes in extra Italian dressing mix into the Pyrex cup; A hand tops the salad with an extra garnish of red bell pepper.

CoE:

<event>Time:00:00-00:10, Des: A close-up shot of a pasta salad is shown, followed by a hand chopping red onions on a wooden cutting board.</event>
 <event>Time:00:10-00:20, Des: The scene transitions to a hand slicing a red bell pepper, then cutting broccoli florets off the stalks.</event>
 <event>Time:00:20-00:30, Des: Next, a can of sliced black olives and a can of artichoke hearts are displayed. A hand slices an artichoke heart in half.</event>
 <event>Time:00:30-00:40, Des: The video then shows a pot of boiling water on a stove, with salt being added. Broccoli is added to the pot three minutes before the pasta finishes cooking.</event>
 <event>Time:00:40-00:50, Des: The pasta is drained and cooled under running water. The cooled pasta is then mixed with the chopped vegetables and other ingredients.</event>
 <event>Time:00:50-01:00, Des: The creamy dressing is poured over the bowl of pasta salad and mixed thoroughly. Finally, a block of feta cheese is crumbled and added to the pasta salad.</event>
 <think>In the video, the preparation of the dressing involves mixing red wine vinegar, water, and Italian dressing mix in a Pyrex cup. The next logical step would be to blend these ingredients together, possibly adding vegetable oil for emulsification. Following this, the dressing should be poured over the salad and mixed thoroughly. The final step mentioned in the question is the addition of feta cheese. The correct sequence aligns with the provided options where the dressing is blended and then mixed with the salad, followed by the addition of feta cheese.</think>
 <answer>C</answer>

Figure 5. An example of CoE paradigm.

Input Video

Question:
 “Based on the given video, predict future events and fill in the potential event in the following sequence: 1. [?] 2. The final scene shows the person in a warm, softly lit room, still mimicking the animated character with red hair, glasses, and a striped shirt. Which of the following best describes the missing event that likely occurred between the currently observed scenes and the final scene?”

Options:

- A. The person mimics a blonde animated character with curly hair, maintaining the same expression in the warm room.
- B. The person mimics an animated white cat wearing a pink bow while staying in the warm, softly lit room.
- C. The person mimics an animated character with a blue suit and a red bow tie, similar to earlier scenes.
- D. The person mimics an animated character with purple hair and a red gem on the forehead while the background changes to a cool, dim room.

CoE:

<event>Time:00:00-00:05, Des: “Challenge Accepted” text appears on screen.</event>
 <event>Time:00:05-00:10, Des: A woman in a white shirt starts mimicking an animated character with a blue suit and a red bow tie.</event>
 <event>Time:00:10-00:15, Des: The woman continues to mimic different animated characters, including one with red hair and glasses, and another with purple hair and a red gem on the forehead.</event>
 <event>Time:00:15-00:20, Des: The woman is now in a warm, softly lit room, continuing to mimic the animated character with red hair, glasses, and a striped shirt.</event>
 <think>Given the sequence of events, it is clear that the woman is transitioning through various animated characters. The final scene shows her in a warm, softly lit room, which suggests a continuation of the previous scenes but with a slight change in the character being mimicked. This suggests that the missing future event should involve continuing to mimic a character, and that this character should not have appeared previously. The last event among the future events indicates that there is no change in the scene within the video, so the future event should involve this person mimicking different characters in the same scene. </think>
 <answer>B</answer>

Figure 6. An example of CoE paradigm.

Input Video

Question:
 “Based on the given video, predict future events and fill in the potential events in the given future events: 1. [?] 2. [?] 3. [?] 4. [A promotional scene appears featuring a black background, an image of the baby (in a blue onesie) in front of a toilet, and call-to-action text urging viewers to check out the previous video]. Which option best represents the correct sequence of future events that bridge the observed scenes and the final scene?”

Options:

A. 1. The video cuts to an indoor cozy scene with a woman on a couch playing with the baby; 2. A quick flash of a promotional black screen with social media web addresses; 3. A return to outdoor walking along a grassy path by the river; 4. Final promotional screen with the baby sitting in front of a toilet and call-to-action text.

B. 1. An outdoor sequence where the baby and adult are seen exploring a dirt path with a visible staircase; 2. A sudden switch to a warm indoor scene with the baby interacting with toys on a colorful blanket; 3. A black screen appears with text promoting social media channels; 4. Final promotional screen with the described baby image and text.

C. 1. A black screen appears with bold red text reading 'SUBSCRIBE YO!' followed by a screen displaying web addresses for social media; 2. The scene returns outdoors near the river showing the adult and baby walking along a grassy path surrounded by trees; 3. The video transitions indoors where a woman is seen on a couch with the baby (now in a white onesie with green text) amid cozy details like a colorful blanket and an open laptop; 4. Final promotional screen featuring a black background with white text urging viewers to 'CHECK OUT MY LAST RYDEN VIDEO!!' alongside an image of the baby in a blue onesie in front of a toilet.

D. 1. A scene shows the baby back on the high chair exploring its tray in a neutral room; 2. The video shifts to an alternate outdoor setting with the adult in a red shirt pointing towards a river; 3. A black screen with red text appears but without any social media details; 4. Final promotional screen appears with minimal text and no clear baby image.

CoE:
 <event>Time:00:00-00:30, Des: A baby in a striped onesie is lying on a quilted blanket, moving its arms and hands playfully.</event>
 <event>Time:00:30-00:45, Des: The baby is now being held by an adult wearing a red hat and a striped shirt, both smiling and engaging with the baby.</event>
 <event>Time:00:45-01:00, Des: The baby is seated in a high chair, smiling and looking around, appearing content and curious.</event>
 <event>Time:01:00-00:30, Des: The baby is being held by an adult outdoors near a river, with the adult talking and gesturing while holding the baby close.</event>
 <event>Time:01:30-01:45, Des: The baby is back in the high chair, exploring its tray, indicating a return to a more structured environment.</event>
 <think>The video progresses from indoor scenes with the baby on a quilted blanket and in a high chair, to an outdoor scene with the baby being held by an adult near a river. The sequence then returns to an indoor setting before concluding with a promotional scene. The options provided should reflect this progression, starting with an outdoor scene, followed by an indoor scene, and ending with a promotional screen.</think>
 <answer>C</answer>

Figure 7. An example of CoE paradigm.

Input Video

Question:
 “Based on the given video, predict future events and fill in the potential events in the following sequence: 1. The video shifts to a party setting with streamers, balloons, and people holding red cups in a dim space lit with pink and blue hues. 2. [?] 3. Soon after, the woman in the black and white striped top (now paired with a white skirt and a large white bow) takes the center stage and begins dancing energetically as a man watches. 4. [?] 5. Later, a woman wearing a black sweater with a patterned collar is shown observing the dancing couple. Which option correctly fills in the two missing events with scenes that occur between the observed scenes and the final end screen?”

Options:

A. After the party is introduced, the dancing couple holds hands and moves in perfect sync before the energetic solo dance, and later the scene transitions to the original indoor setting where the woman is seen speaking expressively.

B. Right after the party is introduced, the indoor set reappears with the woman speaking, then later the dancing couple holds hands and moves in sync before the observer is shown.

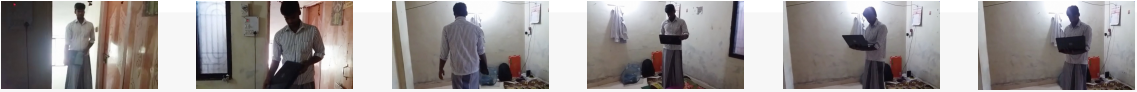
C. Immediately following the party scene, another group of dancers appears performing similar moves, and later, a brief close-up of red cups and streamers is shown before the observer appears.

D. After the party setting is established, a montage of previously seen alley dance moves is replayed, and then a quick scene shows party guests clapping before the observer appears.

CoE:
 <event>Time:00:00-00:28, Des: A woman in a black and white striped top speaks directly to the camera in an indoor setting.</event>
 <event>Time:00:28-00:57, Des: The same woman is now outside in an alleyway, demonstrating various dance moves labeled on-screen.</event>
 <event>Time:00:57-01:16, Des: The woman continues her dance routine in the alleyway, showcasing different moves like 'SpongeBob + Patrick' and 'Cold Shoulder.</event>
 <event>Time:01:16-01:39, Des: The scene returns indoors, where the woman speaks again, this time with a more animated expression.</event>
 <event>Time:01:39-02:00, Des: The video cuts back to the alleyway, where the woman dances alongside a man, performing synchronized moves.</event>
 <event>Time:02:00-02:15, Des: The scene shifts to a party setting with streamers, balloons, and people holding red cups in a dimly lit room.</event>
 <event>Time:02:15-02:30, Des: The woman in the black and white striped top reappears, now in a white skirt and a large white bow, dancing energetically while a man watches.</event>
 <event>Time:2:30-2:45, Des: The video ends with a woman in a black sweater with a patterned collar observing the dancing couple.</event>
 <think> Given the sequence of events, we need to identify the correct option that fits logically between the party scene and the final scene with the observer. Option A correctly transitions from the party scene to the indoor setting where the woman speaks, then to the alleyway dance, and finally back to the original indoor setting where she speaks again. This sequence matches the observed video flow most accurately.</think>
 <answer>A</answer>

Figure 8. An example of CoE paradigm.

Input Video



Question:
 “Based on the given video, predict future events and fill in the potential events in the following future event sequence: 1. [?] 2. [?] 3. [?] 4. The man stands near the wall with hanging clothes under a distinct light source, intently reading or typing on his laptop with a contemplative look. Which sequence of intermediate events best fits the evolution from the observed scenes to this final future scene?”

Options:

- A. 1. The man abruptly leaves the room with the laptop, 2. He returns outdoors in broad daylight, 3. He interacts with multiple strangers, then 4. The man stands near the wall with hanging clothes under a distinct light source, intently reading or typing on his laptop with a contemplative look.
- B. 1. The man remains stationary in the room, 2. He sits down on the colorful mat and starts writing notes by hand, 3. He gradually closes his laptop as the room darkens further, then 4. The man stands near the wall with hanging clothes under a distinct light source, intently reading or typing on his laptop with a contemplative look.
- C. 1. The man moves toward a doorway bathed in natural light and interacts again with the switchboard, 2. He walks around the room pausing by the hanging clothes, 3. He returns to the center of the room, still focused on his laptop with subtle changes in lighting, then 4. The man stands near the wall with hanging clothes under a distinct light source, intently reading or typing on his laptop with a contemplative look.
- D. 1. The man disconnects his laptop and inspects the red bag on the floor, 2. He rearranges the clothes hanging on the wall in a frenzied manner, 3. He sits in silence and looks out of a window for a long pause, then 4. The man stands near the wall with hanging clothes under a distinct light source, intently reading or typing on his laptop with a contemplative look.

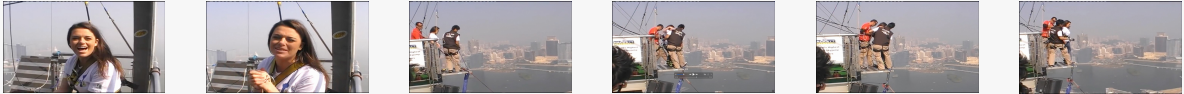
CoE:

```

<event>Time:00:00-0:5, Des: A man enters a room carrying a laptop.</event>
<event>Time:0:5-1:5, Des: The man places the laptop on a colorful mat and begins using it.</event>
<event>Time:1:5-3:0, Des: The man stands up and continues working on the laptop while occasionally looking around the room.</event>
<event>Time:3:0-4:0, Des: The man shifts his position slightly but remains engaged with the laptop, maintaining focus on the screen.</event>
<event>Time:01:39-02:00, Des: The video cuts back to the alleyway, where the woman dances alongside a man, performing synchronized moves.</event>
<event>Time:02:00-02:15, Des: The scene shifts to a party setting with streamers, balloons, and people holding red cups in a dimly lit room.</event>
<event>Time:02:15-02:30, Des: The woman in the black and white striped top reappears, now in a white skirt and a large white bow, dancing energetically while a man watches.</event>
<event>Time:2:30-2:45, Des: The video ends with a woman in a black sweater with a patterned collar observing the dancing couple.</event>
<think> The video shows a man entering a room with a laptop, placing it on a mat, and starting to use it. He then stands up and continues working, occasionally looking around the room. The final scene describes him standing near a wall with hanging clothes, illuminated by a distinct light source, and focused on his laptop. This sequence suggests a progression where the man initially works on the laptop in various positions before settling into a more stationary posture near the wall, indicating a shift from movement to a more contemplative state.
Option C aligns with this progression as it shows the man moving towards a well-lit area, interacting with the environment (hanging clothes), and returning to a focused state near the wall, which matches the final scene described in the question.</think>
<answer>C</answer>
  
```

Figure 9. A bad case of CoE paradigm.

Input Video



Question:
 “Based on the given video, predict future events and fill in the potential events in the given future sequence: 1. [?] 2. A close-up of a sign showing 'HACKETT MACAU TOWER' along with 'The World's Highest Bungee and Skyjump 233 meters' and the website 'www.ajhackett.com'. Which sequence of events best fits in slot 1?”

Options:

- A. The person waves at the camera, then leans forward and jumps off the platform. An aerial view follows as the person is seen diving headfirst with the bungee cord visible, landing on an inflatable platform where they are gradually pulled toward the edge by the cord.
- B. The person stays on the platform and engages in a lengthy conversation with the crew, while no jump occurs; the crew then steps away and the scene abruptly cuts to the close-up of the sign.
- C. Instead of a direct jump, the person performs a series of small hops on the platform, interspersed with short sidesteps, and only later does the camera cut to the detailed close-up of the sign.
- D. The person is seen descending slowly using an elevator mechanism attached to the structure, bypassing any free-fall or dynamic bungee jump, and later the scene shifts to the detailed view of the sign.

CoE:

```

<event>Time:00:00-0:5, Des: A person is seen preparing for a bungee jump at the Hackett Macau Tower. They are wearing safety gear and standing on a platform high above the city.</event>
<event>Time:0:5-1:4, Des: The person waves at the camera, indicating readiness for the jump. The camera captures their expression and the surrounding cityscape.</event>
<event>Time:1:4-2:26, Des: The person is secured with a harness and is being assisted by two individuals. They appear to be getting ready for the jump, with the camera focusing on the preparation process.</event>
<event>Time:2:26-5:0, Des: The person is now on the edge of the platform, surrounded by safety equipment. They are being helped by the crew members who are ensuring everything is secure before the jump.</event>
<event>Time:5:0-7:0, Des: The person is now standing alone on the platform, looking down at the city below. The camera captures the anticipation and the height of the jump.</event>
<think> The video shows a sequence of events leading up to a bungee jump. The person is initially seen preparing, then being assisted, and finally standing alone on the platform. The close-up of the sign provides additional context about the location and activity but does not show the jump itself. Therefore, the most logical sequence that fits the given events would include the jump occurring after the preparation phase, which aligns with option A.</think>
<answer>A</answer>
  
```

Figure 10. A bad case of CoE paradigm.

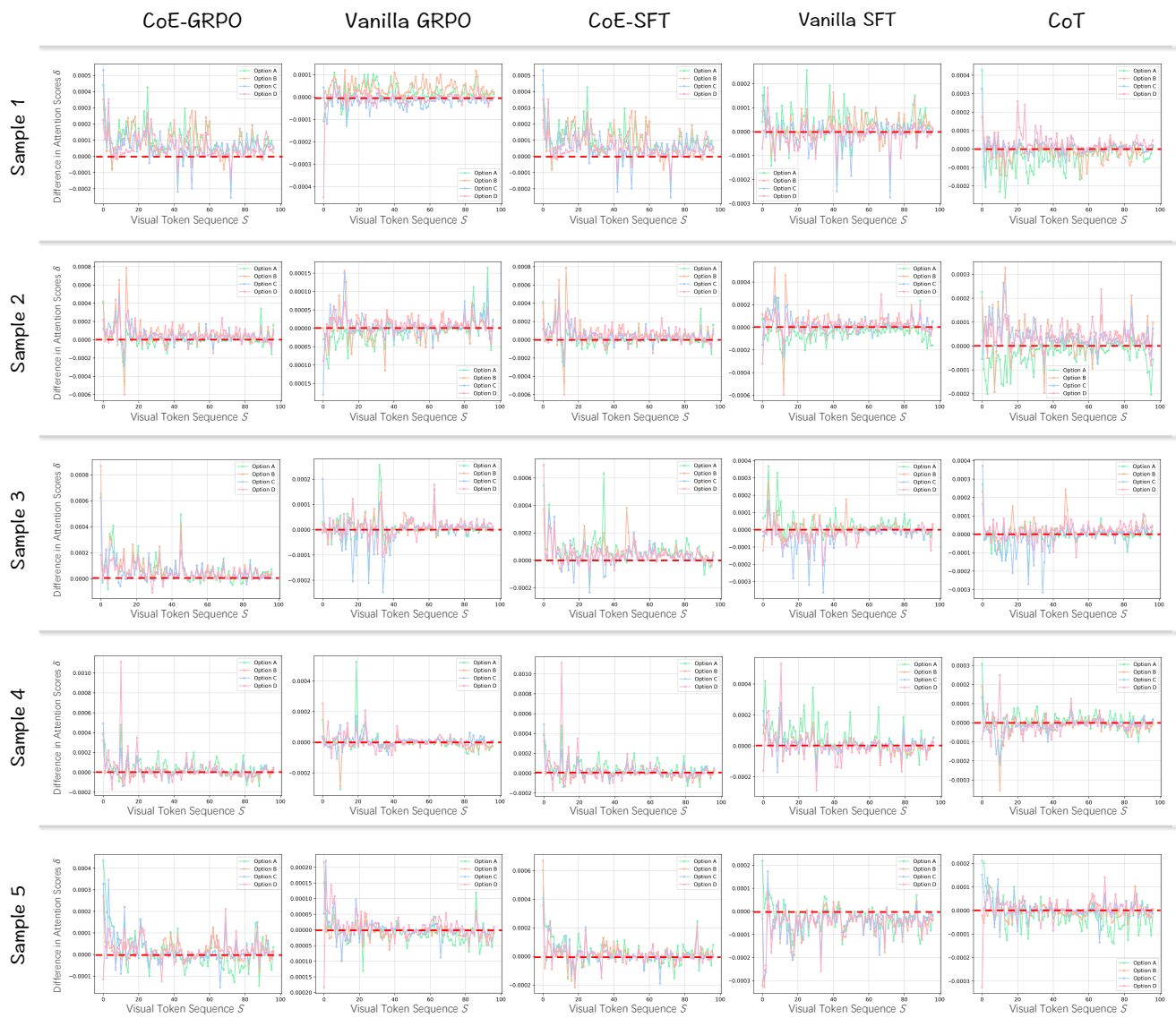


Figure 11. Some examples of attention differences comparing to the vanilla model.

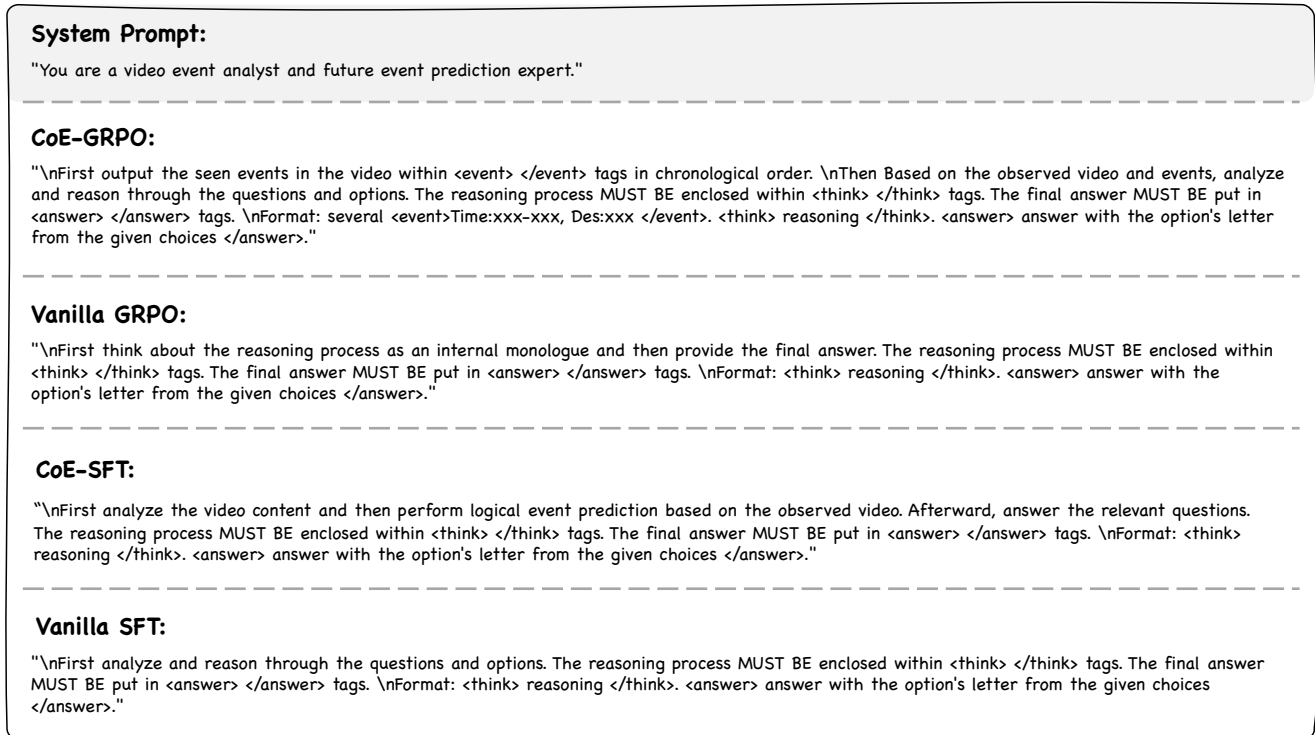


Figure 12. Prompt template for training and inference.

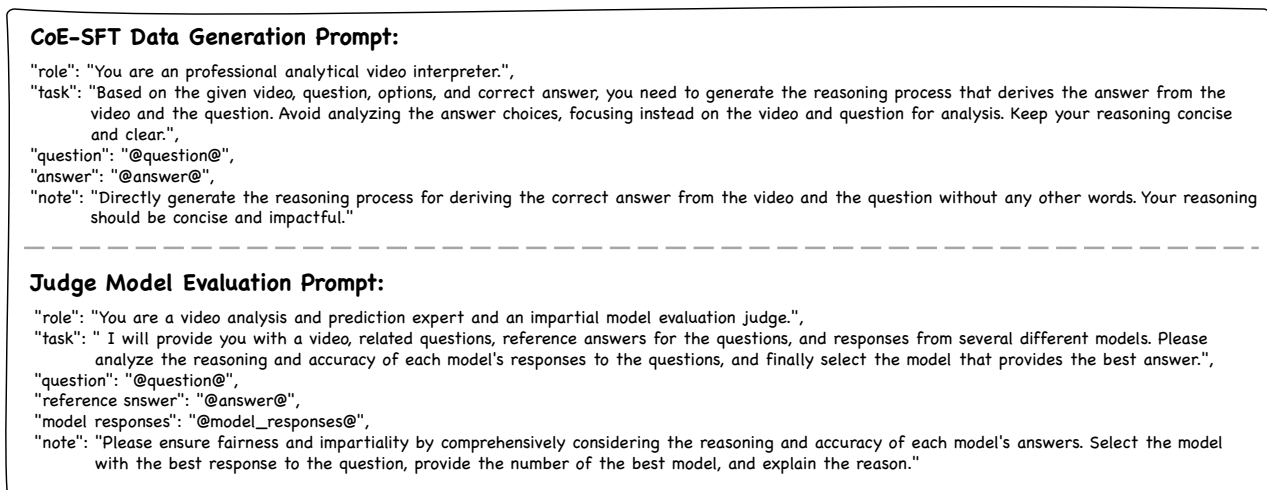


Figure 13. Prompt template for CoE-SFT data generation and judge model evaluation.