AAAI 2026 Supplementary Material

Anonymous submission

Additional Experimental Results

In Causality-aware Cross-modal Hypotheses Selection, we propose a cross-modal causal contrastive learning mechanism to select candidate hypotheses with high causal relevance rather than mere correlation. As shown in Table 1, the performance with CLIP-selected hypotheses is close to random selection, especially on the CIDEr metric. This indicates that CLIP-selected hypotheses contain noticeably fewer key terms than our causal contrastive module, which proofs it fails to capture the causal relationships.

Variant	BLEU@4	METEOR	ROUGE	CIDEr	BERT-S
Random	6.26	13.16	27.58	54.12	36.56
CLIP	6.44	13.28	27.73	54.89	36.69
CCM	**6.54**	**13.41**	**27.95**	**57.04**	**36.80**

Table 1: Ablation study on variants of selection methods in Causality-aware Cross-modal Hypotheses Selection. `CCM` denotes our causal contrastive module.

Additional Qualitative Visualization

In Fig. 1-2, we show more qualitative examples from VAR test and YouCookII test. For VAR test, since the main paper has demonstrated the case where the unknown event H is at the middle of video example $\mathcal{V}$, we include visualization results for the other two cases. For YouCookII test, we present all the three cases. In contrast to the competitors (Liang et al. 2022; Hurst et al. 2024; Wang et al. 2024), plausible inferences for hypotheses are observed for our proposed method, which demonstrates a superior abductive reasoning ability for capturing causal structures among visual events.

Detailed Prompt

Here, we provide detailed prompts for LLM/MLLM usages in the main paper.

Prompt for Video Captioning

In Causality-aware Hypotheses Generation step 1, we use pretrained Qwen2VL-7B (Wang et al. 2024) to generate captions $\mathcal{C}$ for observation events $\mathcal{O}$. The prompt is:

Describe the core events in terms of WHO is doing WHAT, and in WHERE. Do not include irrelevant visual details. The goal is to provide enough factual information to support reasoning and inference. Return the desciption directly without any prefix.

Prompt for Causality-aware Hypotheses Generation

We prompt GPT-4o-mini (Hurst et al. 2024) to infer possible hypotheses based on $\mathcal{C}$. The prompt is:

*You are an event-completion expert. The user inputs a list of detailed event descriptions [str_1 str_2 [MASK]... str_n]. Read the detailed video caption, infer the most plausible event at the '[MASK]' position. Output should be one sentence that includes the subject, action, and location, without lists or extra explanations. Do **not** mention '[MASK]' or refer to this prompt.*

Prompt for Negative Sample Generation

In Causality-aware Hypotheses Generation step 2, we provide GPT-4o-mini (Hurst et al. 2024) with $\mathcal{C}$ and the positive explanation E_h to generate multiple negative samples. The prompt is:

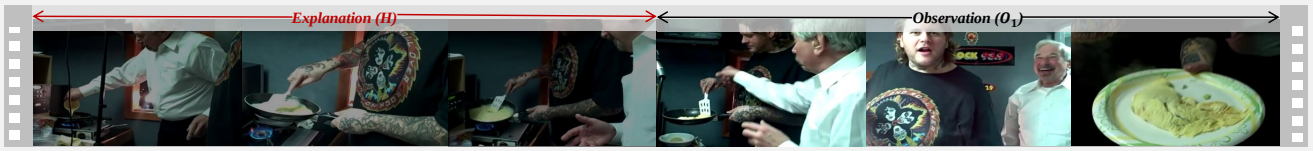
There is a contrastive learning task aimed at matching the missing video description using a series of given videos. The user will input a dict as {'Video descriptions': [str_1 str_2 [MASK] ... str_n], 'Positive sample': 'xxx'}, where '[MASK]' in 'Video descriptions' represents the missing video description, 'Positive sample' is the ground truth description of the missing video. Based on this information, generate a negative sample for the task. The negative sample should differ in semantics from the positive sample, but still align with the logical context of the video descriptions. Format your answer to this template: The negative sample is that <The negative sample of video description>.

Prompt for Abductive Reasoning

Here, we provide prompts for training/inference on the two datasets.

Prompt for VAR dataset. We first provide the basic prompt (w/o generated hypotheses) for training/inference on VAR dataset as below:

SYSTEM: You are an AI assistant able to analyze and infer the most likely explanation for an incomplete set of observation videos.



Groundtruth: [An older man is seen pouring liquid into a pan with another man stirring around the mixture.] [The men help each other cook while laughing and smiling to the camera and end with an egg on a plate and the man holding up a hot plate.]
REASONER: A man is seen speaking to the camera and leads into him trying to solve a rubix cube.
GPT-4o-mini: An abstract visual noise or static display was showcased, which may represent an interruption or a unique artistic representation.
Baseline (Qwen2VL−7B^{FT}): A man is seen speaking to the camera while holding a knife and plate and leads into him cutting up food.
AbductiveMLLM (Ours): A man is seen cooking food on a stove while another man stands next to him.

(a) The case where H is at the beginning of $\mathcal{V}$.

Groundtruth: [A wrestler is seen walking out in front of an audience and sitting on the floor.] [Two men are then seen walking out onto a pit with one throwing dirt and kneeling down.] [The men begin fighting with one pushing another out.]
REASONER: The crowd is seen holding up to the crowd and the audience claps.
GPT-4o-mini: A match was about to take place in a sumo wrestling tournament, with the competitors making their way to the ring and preparing for the bout.
Baseline (Qwen2VL−7B^{FT}): The two sumo wrestlers begin fighting.
AbductiveMLLM (Ours): The two men begin wrestling and one of them falls out of the ring.

(b) The case where H is at the end of $\mathcal{V}$.

Figure 1: Additional qualitative comparison of **`AbductiveMLLM`** on examples from VAR `test`.

USER: The red numbers on each frame represent the event index, with identical numbers for frames in the same event. What most likely happened in event {h_id}? Format your answer to this template: The most likely event is that <The event description>.
{h_id} corresponds to our Number-Prompt on each frame, which denotes the index of H in $\mathcal{V}$. Then we provide the prompt for introducing top-k hypotheses from Causality-aware Hypotheses Generation. We simply add the follow prompt after the USER prompt above:

You may refer to the following information if necessary. Here are some top-ranking inferences: 1-{Top-1 hypothesis} 2-{Top-2 hypothesis} 3-{Top-3 hypothesis}.

Prompt for YouCookII dataset. The primary distinction between YouCookII and VAR is that the ground truths in YouCookII are actions, which are imperative sentences rather than event descriptions. By shifting the reasoning objective from events to actions, we can easily transition from VAR to YouCookII. The SYSTEM prompt is the same as above, and the USER prompt is as below:

*The red numbers on each frame represent the event index, with identical numbers for frames in the same event. What is the most likely Action in event {h_id}? Format your answer to this template: The most likely action is to <The action description>. **<The action description>should begin with a verb (imperative sentence).**

References

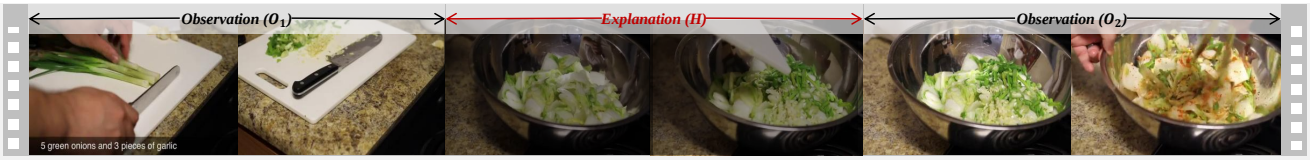
Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Liang, C.; Wang, W.; Zhou, T.; and Yang, Y. 2022. Visual abductive reasoning. In *CVPR*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Groundtruth: [Clean the prawns leaving the tail piece and make 3 small slits on each prawn.] [Take the ready shrimps in a bowl, add sauce and mix them all.] [Take marinated shrimp by tail dip it in batter and coat with bread crumbs and deep fry till golden brown.]
REASONER: Cut the chicken into slices.
GPT-4o-mini: Prepare fried shrimp.
Baseline (Qwen2VL−7B^{FT}): Cut shrimp in half.
AbductiveMLLM (Ours): Peel and devein the shrimp and place in a bowl.

(a) The case where H is at the beginning of $\mathcal{V}$.



Groundtruth: [Chop up the green onion and garlic.] [Add the green onion and garlic to the cabbage.] [Add red pepper and mix.]
REASONER: Add some chopped green onions and chopped green onions.
GPT-4o-mini: Mix salt in well.
Baseline (Qwen2VL−7B^{FT}): Chop the radish.
AbductiveMLLM (Ours): Add the chopped green onions and garlic to the cabbage.

(b) The case where H is at the middle of $\mathcal{V}$.

Groundtruth: [Crush the cabbage in the jar.] [Put cabbage on top and push down.] [Seal the jar.]
REASONER: Add the egg and the egg mixture to the bowl and mix
GPT-4o-mini: Prepare for fermentation.
Baseline (Qwen2VL−7B^{FT}): Put the jar in a cool place.
AbductiveMLLM (Ours): Seal the jar.

(c) The case where H is at the end of $\mathcal{V}$.

Figure 2: Additional qualitative comparison of **`AbductiveMLLM`** on examples from YouCookII `test`.