A. Video Collection

To curate our SmartHome-Bench dataset, we collect videos from public sources, such as YouTube. We craft a keyword set to crawl and identify videos with anomalies in smart homes. To achieve this, we survey the literature on different aspects, such as home security [8], family care [57], and pet monitoring [16]. Additionally, we develop a separate keyword set to capture typical, normal events in smart homes. These keywords are then refined with input from smart home experts. Table 7 shows examples of keywords used in the search process. For each keyword, we collect approximately 20 videos from YouTube, resulting in an initial pool of 8,611 videos. We then filter out irrelevant footage, such as edited content and videos not captured by smart home cameras. For relevant videos that contain advertisements, we trim these segments to ensure the videos are clean. This curation process results in the final SmartHome-Bench dataset, comprising 1,203 videos recorded by both indoor and outdoor smart home cameras.

Table 7. Example keywords for searching normal and abnormal videos.

Type	Example Keywords
Normal Videos	cat playing home cam, squirrels in the yard, rabbits outdoors, dog playtime indoors, baby sleeping crib, kid surveillance, elderly resting safe, senior camera monitoring, senior walking, visitor arrival video, vehicle arriving home, scheduled delivery home, delivery pickup, trees moving backyard, normal weather events, background motion video
Abnormal Videos	pet vomiting home cam, child wandering outside, kid sharp objects, child sudden fall, senior unexpected fall, senior physical distress, elderly rough caregiver, unauthorized entry attempt, package theft, car theft driveway, broken window home, suspicious person home, severe weather property, fire damage home, earthquake home safety, severe wind backyard, thunderstorm backyard, flood property risk

B. Smart Home Anomaly Taxonomy

We present a comprehensive taxonomy for video anomalies in the smart home domain. This taxonomy is developed based on user study, focusing on seven areas like security, senior care, and pet monitoring, and is further refined by smart home experts. Each category is further divided into normal and abnormal videos, with detailed descriptions provided for both.

1. **Wildlife**
   Normal Videos:
   - **Harmless Wildlife:** Harmless wildlife sightings, such as squirrels, birds, or rabbits, moving through the yard.
   - **Common Pests:** Common pest activity that doesn't pose immediate danger (e.g., bugs in the garden).

   Abnormal Videos:
   - **Dangerous Wildlife:** Presence of dangerous wildlife like snakes, spiders, or raccoons that may pose a health risk.
   - **Wildlife Damage:** Any wildlife activity that causes or potentially causes damage to property or threatens human or pet safety.
   - **Indoor Wildlife:** Any wildlife (dangerous or not) that enters a home without clear containment.
2. **Pet Monitoring**
   Normal Videos:
   - **Routine Pet Activity:** Pets engaging in regular play, resting or moving around within designated safe areas.
   - **Safe Interaction:** Pets interacting with known family members or other pets.
   - **Supervised Pets:** Pets accompanied by their guardian without interacting with property or people in harmful ways.

   Abnormal Videos:
   - **Unattended Pets:** Pets left outside alone for extended periods.
   - **Escape Attempts:** Pets attempting to escape, leaving the designated area, or exhibiting behaviors indicating escape attempts.
   - **Destructive Behavior:** Pets causing property damage by actions like chewing, scratching, or digging.
   - **Distress Signals:** Behaviors that indicate illness or distress, like vomiting, excessive scratching, or erratic movements.
   - **Conflict or Injury Risk:** Any interaction with others that could lead to conflict or injury.
3. **Baby Monitoring**
   Normal Videos:

- **Safe Play:** Baby engaging in play or sleep within safe zones or under supervision.
- **Caregiver Interaction:** Harmless interactions between the baby and caregivers.

Abnormal Videos:

- **Near Danger:** Baby nearing dangerous zones (e.g., staircases, swimming pools) without adult supervision.
- **Unattended Baby:** Baby wandering outside a crib, stroller, or designated play area without adult presence.
- **Injury Risk:** Sudden, unexpected falls that may lead to injury.
- **Baby Abuse:** Any abusive behavior toward the baby, such as hitting, or forcing them to act against their will.

4. **Kid Monitoring**

Normal Videos:

- **Safe Play:** Kids playing or moving around indoors or outdoors within designated areas.
- **Routine Activities:** Regular daily activities under adult supervision.

Abnormal Videos:

- **Wandering:** Kids found wandering outdoors or in dangerous locations without adult supervision.
- **Dangerous Actions:** Dangerous actions indoors (e.g., playing with sharp objects, accessing restricted areas) or significant health/safety concerns (e.g., choking hazards).
- **Injury Risk:** Sudden, unexpected falls that may lead to injury.

5. **Senior Care**

Normal Videos:

- **Routine Activity:** Seniors engaging in routine activities like walking, resting, or interacting with caregivers or family.

Abnormal Videos:

- **Senior Falls:** Sudden, unexpected falls that may lead to injury.
- **Distress Signals:** Signs of distress or calls for help through hand gestures or unusual body language.
- **Elder Abuse:** Any abusive or rough behavior by caregivers toward seniors, including verbal and physical abuse.

6. **Security**

Normal Videos:

- **Routine Activity:** Routine activity of homeowners, known visitors, or vehicles arriving and leaving.
- **Scheduled Delivery:** Scheduled package deliveries or pickups without interference.

Abnormal Videos:

- **Unauthorized Entry:** Motion or presence indicating potential break-ins, or trespassing.
- **Suspicious Loitering:** Loitering individuals or those wearing unusual attire that deviates from the norm.
- **Forced Entry:** Forced entry attempts, such as fiddling with locks, tampering with doors or windows, or trying to enter a home or vehicle through unconventional means.
- **Theft or Vandalism:** Unauthorized removal of packages, vehicles, or other items.
- **Property Damage:** Acts of property damage like graffiti, broken windows, car crashes, or other forms of vandalism.
- **Violence or Threats:** Actions that might cause harm, such as kidnapping, aggressive confrontations, or any threatening behavior.
- **Disturbing Behavior:** Unusual or eccentric behavior by individuals that could alarm or frighten viewers.

7. **Other Category**

Normal Videos:

- **Everyday Activity:** Videos that do not fit any of the above categories but show harmless, everyday activities, such as trees waving, normal weather events, or background motion.

Abnormal Videos:

- **Severe Weather:** Severe weather conditions or natural disasters like fires, earthquakes, floods, or storms causing property damage or safety hazards.
- **Unexplained Phenomena:** Unexplained phenomena of inanimate objects.
- **Falling Objects:** Sudden, unexpected falls of inanimate objects that may cause damage or injury.
- **Risky activities:** Irregular activities that do not fit into other categories but may pose risks or concerns.

C. Video Annotation

During the video annotation process, we assign unique IDs to the downloaded videos to prevent annotators from being influenced by the original titles or metadata. The annotators classify each video into one or more of the seven categories in the taxonomy outlined in Appendix B, as real-world events in a single video may span multiple categories. Each video is then assigned an anomaly tag of `normal`, `abnormal`, or `vague abnormal`, based on the definitions outlined in the

taxonomy. The `vague abnormal` category is created for videos where annotators cannot reach a consensus on whether the content is normal or abnormal. This category is specifically introduced to challenge the video anomaly detection (VAD) capabilities of multi-modal large language models (MLLMs) with videos that are difficult for even humans to classify. A `vague normal` category is not included, as any ambiguity regarding the presence of an anomaly is classified under `vague abnormal`.

We instruct annotators to write high-quality video descriptions and provide detailed reasoning for the assignment of each video's anomaly tag. These annotations establish a strong foundation for future research by enabling the generation of diverse question-answer pairs to assess the video understanding and reasoning capabilities of MLLMs. Additionally, the inclusion of ground-truth reasoning ensure a transparent inference process for classifying normal and abnormal videos, which can be leveraged to fine-tune MLLMs and improve anomaly detection accuracy in smart home scenarios. To maintain consistency and quality across video descriptions and reasoning annotations, we use the Gemini-1.5-pro model to generate initial drafts. Annotators then review each video and refine or rewrite these drafts according to three main criteria: (1) clarity and precision of language, (2) alignment of descriptions and reasoning with the video content, and (3) accuracy in identifying key elements such as objects triggering anomalies, abnormal movements, participants, and environmental conditions.

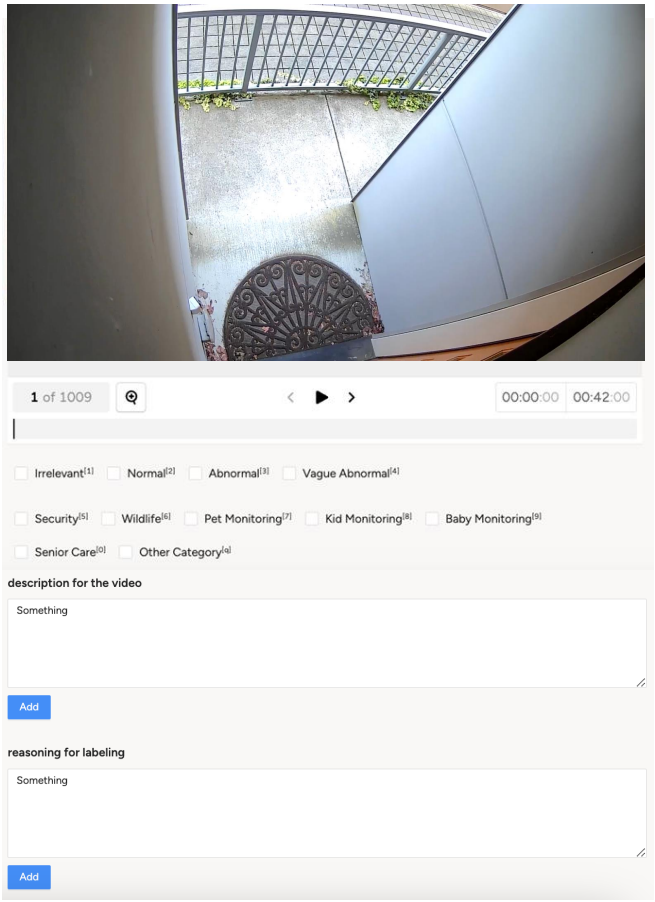

Figure 9. The UI enables annotators to label videos by selecting event categories, assigning anomaly tags, and providing detailed video descriptions along with the reasoning behind the observed anomalies or normality.

To streamline the annotation process and maximize efficiency, annotators use a customized user interface (UI), shown in Figure 9, to label each video's event category and anomaly tag, as well as to manually write the description and reasoning. To ensure the quality and consistency of the annotations, we conduct a human review of a randomly select 200 videos after the initial round of annotation .

Following the annotation process for all 1,203 videos, the statistics of the SmartHome-Bench dataset are presented in Figure 1a of the main paper. The dataset shows a balanced distribution between abnormal and normal videos, with the

security category containing the largest number of videos among the seven event categories. Additionally, Figure 10 illustrates the distribution of video durations and word counts for descriptions and reasoning annotations. The average video length is approximately 20 seconds, with most clips being shorter than 80 seconds. This duration aligns well with the frame-processing limitations of some existing MLLMs, enabling relatively comprehensive predictions in VAD tasks. The word count distribution reveals that reasoning annotations are typically more concise than descriptions, as they focus solely on the key event leading to the assigned anomaly tag. In contrast, descriptions provide a detailed account of all events within the video.

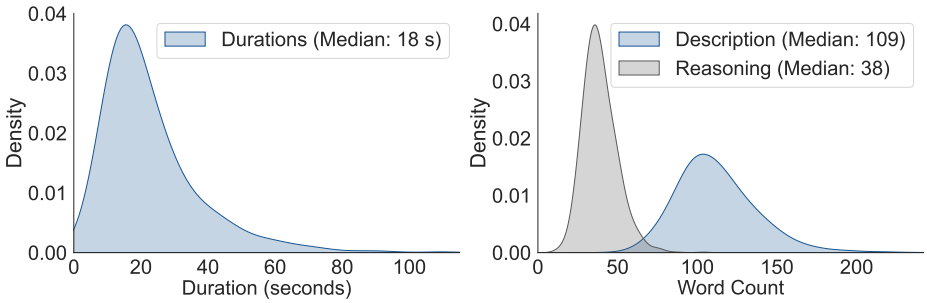

Figure 10. Distribution of video durations and word counts for human-annotated video descriptions and reasoning.

D. Prompts for Adaptation Methods and In-Depth Analysis

We provide all prompts used for adaptation methods and error diagnosis in in-depth analysis as follows.

D.1. System Prompt for Vanilla Adaptations

Figure 11 shows the prompts used in zero-shot prompting for the VAD task.

Zero-Shot Prompting for Video Anomaly Detection

Please watch the video carefully and determine whether it contains any anomalies.

Response Format:

Reply using the following format:

```
{
    "anomaly": 0 or 1  // 0 for no anomaly detected, 1 for anomaly detected
}
```

Figure 11. System prompts adopted in zero-shot prompting for VAD. MLLMs are prompted directly to return a binary anomaly label.

Figure 12 shows the prompts used in chain-of-thought (CoT) prompting for the VAD task.

Chain-of-Thought Prompting for Video Anomaly Detection

You are an excellent smart video vigilance expert agent in the smart home security domain.

Task Instruction:

You are given a smart home video clip, and your job is to carefully identify potentially risky, suspicious, or anomalous situations captured by the surveillance cameras. These cameras are set up by users to enhance their safety and security. Keep in mind that the people in the video may or may not be the camera owners.

Anomaly Definition:

In this context, anomalies refer to behaviors or events that raise concerns related to security, personal safety, child safety, wildlife alerts, unusual pet behavior, senior monitoring, or any other situations that seem out of the ordinary.

Response Format:

Please think step by step and respond using the format below:

```
{
    "video_description": "A concise description of the video content, including objects,
    movements, and environmental conditions (max 200 words)",
    "reasoning": "Detailed reasoning for why the situation is considered abnormal or
    concerning, if applicable (max 100 words)",
    "anomaly": 0 or 1  // 0 for no anomaly detected, 1 for anomaly detected
}
```

Figure 12. System prompts adopted in CoT prompting include task instructions, smart home anomaly definitions, and video input, guiding MLLMs to complete the task in three steps: generating video descriptions, providing reasoning, and predicting the anomaly label.

Figure 13 shows the prompts used in the few-shot CoT prompting for the VAD task.

Few-Shot Chain-of-Thought Prompting for Video Anomaly Detection

You are an excellent smart video vigilance expert agent in the smart home security domain.

Task Instruction:

You are given a smart home video clip, and your job is to carefully identify potentially risky, suspicious, or anomalous situations captured by the surveillance cameras. These cameras are set up by users to enhance their safety and security. Keep in mind that the people in the video may or may not be the camera owners.

Anomaly Definition:

In this context, anomalies refer to behaviors or events that raise concerns related to security, personal safety, child safety, wildlife alerts, unusual pet behavior, senior monitoring, or any other situations that seem out of the ordinary.

Response Format:

Please think step by step and respond using the format below:

```
{
  "video_description": "A concise description of the video content, including objects, movements, and environmental conditions (max 200 words)",
  "reasoning": "Detailed reasoning for why the situation is considered abnormal or concerning, if applicable (max 100 words)",
  "anomaly": 0 or 1  // 0 for no anomaly detected, 1 for anomaly detected
}
```

Reference Examples:

1. Example 1:

```
{
  "video_description": "The video shows a young child running towards a swimming pool. The child jumps into the pool and does not resurface. A man, who appears to be a neighbor, jumps the fence and pulls the child from the pool. The man then performs CPR on the child.",
  "reasoning": "The situation is extremely concerning as a young child jumped into a pool and was unable to resurface. This is a life-threatening situation and requires immediate intervention, which thankfully the neighbor provided.",
  "anomaly": 1
}
```

2. Example 2:

```
{
  "video_description": "A security camera captures a woman attempting to wrangle a small dog on a leash in a driveway. The dog breaks free and runs off-screen. The woman briefly chases after the dog before giving up and returning to the house.",
  "reasoning": "The dog escaping its leash could be concerning for the owner as it raises concerns about the dog's safety and well-being. The dog running loose could lead to it getting lost or potentially encountering dangerous situations.",
  "anomaly": 1
}
```

3. Example 3:

```
{
  "video_description": "The video, taken from a smart doorbell camera, shows a man attempting to break into a house. He is using a crowbar to pry open the front door. The man is wearing a black shirt, black shorts, a maroon beanie, and blue gloves.",
  "reasoning": "The man's actions are clearly suspicious and illegal. Attempting to force entry into a home using a crowbar is a crime. His behavior suggests criminal intent, posing a serious security threat.",
  "anomaly": 1
}
```

Figure 13. System prompts adopted in few-shot CoT prompting for VAD. Each example provided includes a video description, anomaly reasoning, and the corresponding ground-truth anomaly label.

D.2. System Prompt for In-Context Learning

The prompts used in in-context learning (ICL) for the VAD task are shown in Figure 14.

In-Context Learning Prompting for Video Anomaly Detection

You are an expert in smart video surveillance with a focus on smart home security.

Task Guidelines:

You are given a smart home video clip and a set of rules (taxonomy) for identifying anomalies in various smart home scenarios. Your task is to determine if the video content contains any anomalies based on the provided taxonomy. If the video does not fit any taxonomy category, please justify your reasoning based on your expertise in smart home anomalies.

Anomaly Taxonomy:

1. Security
   - Normal Videos:
     - Routine activity of homeowners, known visitors, or vehicles arriving and leaving.
     - Scheduled package deliveries or pickups without interference.
   - Abnormal Videos:
     - Motion or presence indicating potential break-ins or trespassing.
     - Loitering individuals or those wearing unusual attire that deviates from the norm.
     - Forced entry attempts, such as fiddling with locks, tampering with doors or windows, or trying to enter a home or vehicle through unconventional means.
     - Unauthorized removal of packages, vehicles, or other items.
     - Acts of property damage like graffiti, broken windows, car crashes, or other forms of vandalism.
     - Actions that might cause harm, such as kidnapping, aggressive confrontations, or threatening behavior.
     - Unusual or eccentric behavior by individuals that could alarm or frighten viewers.
2. Wildlife

   .....
7. Other Categories
   - Normal Videos:
     - Videos that do not fit any of the above categories but show harmless, everyday activities, such as trees waving, normal weather events, or background motion.
   - Abnormal Videos:
     - Severe weather conditions or natural disasters like fires, earthquakes, floods, or storms causing property damage or safety hazards.
     - Unexplained phenomena of inanimate objects.
     - Sudden, unexpected falls of inanimate objects that may cause damage or injury.
     - Irregular activities that do not fit into other categories but may pose risks or concerns.

Response Format:

Please analyze the video and provide your response in the following format:

```
{
  "video_description": "A concise description of the video content, including objects, movements, and environmental conditions (max 200 words)",
  "reasoning": "Detailed reasoning for why the situation is considered abnormal or concerning, if applicable (max 100 words)",
  "anomaly": 0 // 0 for no anomaly detected, 1 for anomaly detected
}
```

Figure 14. System prompts adopted in ICL for VAD. Building upon the CoT prompt, we include the complete anomaly taxonomy as a reference.

D.3. System Prompt for Taxonomy-Driven Reflective LLM Chain

The prompts used in the taxonomy-driven reflective LLM chain (TRLC) framework for the VAD task are detailed as follows. First, the prompts in Figure 15 are used in step (a) of the TRLC to generate rules from the complete video anomaly taxonomy, with the resulting rules from step (a) shown in Figure 16. Next, the prompts in Figure 17 are employed to predict the initial detection for the VAD task. Finally, the self-reflection step is carried out using the prompts provided in Figure 18.

TRLC for Video Anomaly Detection: Taxonomy-Driven Rule Generation (Step a)

You are a smart home surveillance monitor tasked with deriving general rules for detecting anomalies in video clips. The anomalies are categorized under the following taxonomy:

Anomaly Taxonomy:

```
{
  "Security": {
    "Normal Videos": [
      "Routine activity of homeowners, known visitors, or vehicles arriving and leaving.",
      "Scheduled package deliveries or pickups without interference."
    ],
    "Abnormal Videos": [
      "Motion or presence indicating potential break-ins or trespassing.",
      "Loitering individuals or those wearing unusual attire that deviates from the norm.",
      "Forced entry attempts, such as fiddling with locks, tampering with doors or
      windows, or trying to enter a home or vehicle through unconventional means.",
      "Unauthorized removal of packages, vehicles, or other items.",
      "Acts of property damage like graffiti, broken windows, car crashes, or other forms
      of vandalism.",
      "Actions that might cause harm, such as kidnapping, aggressive confrontations, or
      threatening behavior.",
      "Unusual or eccentric behavior by individuals that could alarm or frighten viewers."
    ]
  },
  ......
  "Other Categories": {
    "Normal Videos": [
      "Videos that do not fit any of the above categories but show harmless, everyday
      activities, such as trees waving, normal weather events, or background motion."
    ],
    "Abnormal Videos": [
      "Severe weather conditions or natural disasters like fires, earthquakes, floods, or
      storms causing property damage or safety hazards.",
      "Unexplained phenomena of inanimate objects.",
      "Sudden, unexpected falls of inanimate objects that may cause damage or injury.",
      "Irregular activities that do not fit into other categories but may pose risks or
      concerns."
    ]
  }
}
```

Guidelines:

1. Ensure each rule is distinct and not repetitive or similar to others.
2. Each rule should be general enough to apply to various scenarios without being overly specific.

Reference Rule:

"Pets seen outdoors without their owner, which could suggest escape or abandonment."

Output Format:

Please structure the output as follows, with each rule being a clear, concise statement and no more than 50 words:

```
{
  "rules": [
    "Rule 1: [Description of rule (max 50 words)]",
    "Rule 2: [Description of rule (max 50 words)]"
    // Add additional rules as needed, each under 50 words
  ]
}
```

Figure 15. System prompts adopted in step (a) of the TRLC for VAD: taxonomy-driven rule generation.

10 Rules Generated in Step a of the TRLC

Generated Rules:

```
{
  "rules": [
    "Rule 1: Detect unusual or eccentric behavior by individuals that could alarm or
    frighten viewers.",
    "Rule 2: Identify pets outdoors without their owner.",
    "Rule 3: Monitor the presence of wild animals near the house, posing potential danger
    or risk of property damage.",
    "Rule 4: Detect individuals attempting unauthorized access to the property or vehicles,
    indicating potential theft.",
    "Rule 5: Monitor severe weather conditions or disasters like fires or earthquakes
    causing property damage or safety hazards.",
    "Rule 6: Identify children outside the home without supervision.",
    "Rule 7: Detect pets causing property damage or attempting to escape.",
    "Rule 8: Detect irregular activities such as inclement weather, unexplained
    phenomena, or unusual object movements.",
    "Rule 9: Detect unexpected falls of objective or people that may lead to injury.",
    "Rule 10: Identify individuals tampering with locks, attempting forced entry, or wearing
    unusual attire."
  ]
}
```

Figure 16. 10 rules generated from the full video anomaly taxonomy in step (a) of TRLC by GPT-4o.

TRLC for Video Anomaly Detection: Initial Prediction (Step b)

You are an excellent smart video vigilance expert agent in the smart home security domain.

Task Instruction:

You are given a smart home video clip, and your job is to carefully identify potentially risky, suspicious, or anomalous situations captured by the surveillance cameras. These cameras are set up by users to enhance their safety and security. Keep in mind that the people in the video may or may not be the camera owners.

Anomaly Definition:

In this context, anomalies refer to behaviors or events that raise concerns related to security, personal safety, child safety, wildlife alerts, unusual pet behavior, senior monitoring, or any other situations that seem out of the ordinary.

Response Format:

Please think step by step and respond using the format below:

```
{
  "video_description": "A concise description of the video content, including objects,
  movements, and environmental conditions (max 200 words)",
  "reasoning": "Detailed reasoning for why the situation is considered abnormal or
  concerning, if applicable (max 100 words)",
  "anomaly": 0 or 1  // 0 for no anomaly detected, 1 for anomaly detected
}
```

Figure 17. System prompts adopted in step (b) of the TRLC for VAD: initial prediction. (These prompts are identical to the CoT prompts shown in Figure 12).

TRLC for Video Anomaly Detection: Self-Reflection (Step c)

You are an advanced smart video surveillance expert in the smart home security domain. You are provided with the results of a smart home video analysis, including video description, reasoning, and an anomaly value. Additionally, you have a set of rules for anomaly detection.

Task Guidelines:

Your task is to review the provided rules. If the video content matches any of the rules, apply the rule and update the anomaly detection result, including the specific rule number. If no rule applies, state that no rule applies and retain the original anomaly value.

Video Anomaly Result:

The video anomaly result is:

```
{
   "video_description": "{video_description}",
   "reasoning": "{reasoning}",
   "anomaly": {anomaly}
}
```

Anomaly Rules:

The rules provided for anomaly detection are:

```
{
   "rules": [
      "Rule 1: Detect unusual or eccentric behavior by individuals that could alarm or
      frighten viewers.",
      "Rule 2: Identify pets outdoors without their owner.",
      "Rule 3: Monitor the presence of wild animals near the house, posing potential danger
      or risk of property damage.",
      "Rule 4: Detect individuals attempting unauthorized access to the property or vehicles,
      indicating potential theft.",
      "Rule 5: Monitor severe weather conditions or disasters like fires or earthquakes
      causing property damage or safety hazards.",
      "Rule 6: Identify children outside the home without supervision.",
      "Rule 7: Detect pets causing property damage or attempting to escape.",
      "Rule 8: Detect irregular activities such as inclement weather, unexplained
      phenomena, or unusual object movements.",
      "Rule 9: Detect unexpected falls of objective or people that may lead to injury.",
      "Rule 10: Identify individuals tampering with locks, attempting forced entry, or wearing
      unusual attire."
   ]
}
```

Response Format:

Please think step-by-step and respond using the format below:

```
{
   "Rule_Reasoning": "If the video matches a rule, provide reasoning based on the specific
   rule number. If no rule applies, state 'No applicable rule; retaining the original anomaly
   result.'",
   "updated_anomaly": 0 or 1  // Based on the rule application, update the anomaly
   detection result: 0 for no anomaly detected, 1 for anomaly detected
}
```

Figure 18. System prompts adopted in step (c) of the TRLC for VAD: self-reflection.

D.4. System Prompt for Error Diagnosis in In-Depth Analysis

We use the prompts in Figure 19 and Figure 20 to evaluate MLLM-generated video descriptions and reasoning against human-annotated counterparts, respectively.

MLLM Description Performance Evaluation for Video Anomaly Detection

You are an expert in smart video surveillance. Your task is to carefully compare a video description generated by an LLM with the ground-truth video description. Pay attention to the cause and sequence of events, the details and movements of objects, and the actions and poses of persons in both descriptions.

Task Guidelines:

Based on your observations, select the most appropriate option(s) from the following list that describe the LLM video description:

1. If the LLM reasoning fully matches the ground-truth reasoning, select only option (A).
2. If the LLM description is 'NAN', select only option (F)
3. If the ground-truth description is 'NAN', select only option (H) and set the Reason to 'The ground-truth description is NAN'.
4. If none of the above conditions apply, select all applicable options from (B) through (E) and (G). Multiple options may be selected

LLM Description:

{llm_descriptions}

Ground-Truth Reasoning:

{true_description}

Question:

What situation describes the video description generated by the LLM?

Options:

A. Successfully Matched the Ground Truth: The LLM video description captures all important details in the ground truth.
B. Misinterpretation of Events or Actions: The LLM incorrectly understands or describes what is happening in the video, such as incorrect motion directions or actions.
C. Omission of Key Details: The LLM misses important events or abnormal occurrences that are crucial to an accurate description of the video.
D. Addition of Non-Existent Content (Hallucinations): The LLM includes events, objects, or details in its description that are not present in the actual video.
E. Lack of Contextual Understanding: The LLM fails to grasp the context, nuances, or underlying meaning of the video due to limited knowledge or understanding; for example, failing to recognize that the person in the video is the homeowner and regarding him as a suspicious stranger.
F. Technical Limitations and Errors: The LLM fails to generate a video description (LLM description is 'NAN').
G. Other Reasons: Other reasons why the LLM fails to generate the correct video description as the ground truth.
H. Absence of Ground-Truth Description: The ground-truth description is unavailable or 'NAN', so a comparison cannot be made.

Response Format:

Please provide your assessment in the following JSON format, without any additional text or commentary:

```
{
  "Option": ["A"], // or ["B", "C"] // List of selected option letters
  "Reason": "Your reason for selecting the option(s); specifically, if you selected (G), please specify the reason that the LLM description is wrong (max 100 words)"
}
```

Figure 19. System prompts adopted in evaluating the MLLM-generated video description for VAD.

MLLM Reasoning Performance Evaluation for Video Anomaly Detection

You are an expert in smart video surveillance. Your task is to compare the reasoning about anomalies in a video generated by an LLM with the ground-truth reasoning. Pay close attention to the causes and sequences of events, the details and movements of objects, and the actions and poses of persons in both descriptions.

Task Guidelines:

Based on your observations, select the most appropriate option(s) from the list below that describe the LLM's reasoning about anomalies in the video:

1. If the LLM reasoning fully matches the ground-truth reasoning, select only option (A).
2. If the LLM reasoning is 'NAN', select only option (F).
3. If the ground-truth reasoning is 'NAN', select only option (H) and set the Reason to 'The ground-truth reasoning is NAN'.
4. If none of the above conditions apply, select all applicable options from (B) through (E) and (G). Multiple options may be selected.

LLM Reasoning:

{llm_explanations}

Ground-Truth Reasoning:

{true_explanation}

Question:

Based on the comparison, which of the following options best describe the LLM's reasoning about anomalies in the video?

Options:

A. Successfully Matched the Ground Truth: The LLM's reasoning captures all important details in the ground truth.
B. Misinterpretation of Events or Actions: The LLM incorrectly understands events or actions in the video, leading to incorrect identification of anomalies.
C. Omission of Key Abnormal Events: The LLM misses significant abnormal occurrences, resulting in incomplete anomaly detection.
D. Addition of Non-Existent Anomalies (Hallucinations): The LLM identifies anomalies that don't exist in the video, generating false positives.
E. Lack of Contextual Understanding: The LLM fails to grasp the full context of the video, misinterpreting normal events as anomalies or overlooking actual anomalies.
F. Technical Limitations and Errors: The LLM fails to generate reasoning about the video (LLM reasoning is 'NAN').
G. Other Reasons: Other reasons why the LLM fails to generate the correct video anomaly reasoning as the ground truth.
H. Absence of Ground-Truth Reasoning: The ground-truth reasoning is unavailable or 'NAN', so a comparison cannot be made.

Response Format:

Please provide your assessment in the following JSON format, without any additional text or commentary:

```
{
  "Option": ["A"],  // or ["B", "C"] // List of selected option letters
  "Reason": "Your reason for selecting the option(s); if you selected (G), please specify the reason that the LLM reasoning is wrong (max 100 words)"
}
```

Figure 20. System prompts adopted in evaluating the MLLM-generated video reasoning for VAD.

Table 8. Performance of MLLMs with two prompt frames: accuracy, precision, recall (%), and processing time (s) compared across different MLLMs using zero-shot prompting (AD: anomaly detection, ND: normality detection).

Model	Accuracy		Precision		Recall		Video Processing Time	
	AD	ND	AD	ND	AD	ND	AD	ND
Gemini-1.5-flash	58.44	72.90	79.22	81.36	31.12	64.56	3.43	3.26
Gemini-1.5-pro	57.36	**74.15**	**84.34**	**86.58**	25.73	61.63	4.14	4.02
GPT-4o	68.41	70.74	80.09	82.07	55.16	58.55	10.15	9.79
GPT-4o-mini	69.91	73.07	76.52	78.66	63.79	**68.72**	10.09	10.39
Claude-3.5-sonnet	**70.82**	74.06	69.66	82.97	**81.36**	65.33	20.87	21.51
VILA-13b	46.05	55.28	0.00	78.46	0.00	23.57	**1.38**	**1.28**

E. Additional Experiments

E.1. Comparison between Anomaly Detection and Normality Detection

Anomaly detection is a classical binary classification task [6]. In the context of VAD, we employ two distinct prompt frames to evaluate the accuracy of MLLMs in this classification task. First, we prompt the MLLMs to identify abnormal events within a sequence of normal activities, targeting the anomaly detection task. Conversely, given that "normal videos" constitute the majority of training data [40], we also frame the task as a normality detection issue, prompting MLLMs to justify whether a video is normal. This bidirectional approach allows for a comprehensive evaluation of the MLLMs' capabilities in understanding and reasoning about smart home video clips, highlighting performance differences across different task frames in MLLM-based VAD.

Zero-Shot Prompting The zero-shot prompt for anomaly detection is illustrated in Figure 11, while the prompt for normality detection is provided in Figure 21. Table 8 presents the VAD results for both anomaly detection and normality detection tasks using zero-shot prompting. All MLLMs, except Claude-3.5-sonnet, achieve higher accuracy, precision, and recall in the normality detection task. VILA-13b classifies all videos as normal when tasked with anomaly detection, emphasizing its limitations in zero-shot VAD tasks, despite being the fastest model in processing videos. Given that VAD is a binary classification task, the random guess accuracy is 50%. Even the best-performing MLLMs achieve accuracy close to this threshold, highlighting their limited understanding of anomalies in smart home contexts. These results likely reflect the models' training on datasets primarily composed of normal videos, leading to stronger prior knowledge of normal events in smart home scenarios.

Zero-Shot Prompting for Video Normality Detection

Please watch the video carefully and determine whether the situation captured in the video is normal.

Response Format:

Reply using the following format:

```
{
   "normality": 0 or 1 // 0 for the video is normal, 1 for anomaly detected
}
```

Figure 21. System prompts adopted in zero-shot prompting for video normality detection.

CoT Prompting Given that all MLLMs perform better on normality detection than anomaly detection with zero-shot prompting, an important question arises: does this trend continue with CoT prompting?

To investigate, we evaluate CoT performance for both anomaly detection and normality detection. The prompts used are detailed in Figure 12 and Figure 22, respectively. As shown in Table 9, for the AD results, CoT prompting improves accuracy and recall compared to the zero-shot prompting in Table 8, meeting expectations for CoT's effectiveness. However, performance in normality detection declines with CoT prompting. While four MLLMs achieve over 90% precision in the

Chain-of-Thought Prompting for Video Normality Detection

You are an excellent smart video vigilance expert agent in the smart home security domain.

Task Instruction:

You are given a smart home video clip, and your job is to carefully identify normal, expected, or ordinary situations captured by the surveillance cameras. These cameras are set up by users to enhance their safety and security. Keep in mind that the people in the video may or may not be the camera owners.

Normality Definition:

In this context, normality refers to behaviors or events that are typical and do not raise concerns related to security, personal safety, child safety, wildlife activity, pet behavior, senior monitoring, or any other situations that seem usual.

Response Format:

Please think step by step and respond using the format below:

```
{
  "video_description": "A concise description of the video content, including objects,
  movements, and environmental conditions (max 200 words)",
  "reasoning": "Detailed reasoning for why the situation is considered normal and not
  concerning, if applicable (max 100 words)",
  "normality": 0 or 1 // 0 for the video is normal, 1 for anomaly detected
}
```

Figure 22. System prompts adopted in CoT prompting for video normality detection.

Table 9. Performance of MLLMs with two prompt frames: accuracy, precision, recall (%), and processing time (s) compared across different MLLMs using CoT prompting (AD: anomaly detection, ND: normality detection).

Model	Accuracy		Precision		Recall		Video Processing Time	
	AD	ND	AD	ND	AD	ND	AD	ND
Gemini-1.5-flash	69.58	45.47	74.44	40.00	66.41	2.16	**4.61**	**4.57**
Gemini-1.5-pro	**74.06**	**61.60**	**83.77**	93.90	64.41	**30.82**	7.05	6.83
GPT-4o	72.57	57.94	83.02	**100.00**	61.79	22.03	12.55	14.27
GPT-4o-mini	68.83	49.46	68.07	**100.00**	**79.51**	6.32	12.28	13.39
Claude-3.5-sonnet	71.90	54.20	83.44	95.37	59.78	15.87	24.49	24.09
VILA-13b	68.41	43.39	68.45	13.64	76.89	0.92	6.74	11.56

normality detection task, the overall accuracy drops significantly compared to the ND results in Table 8. This suggests that while MLLMs have a solid grasp of normality, CoT prompting reinforces their existing strengths without addressing their weaknesses in anomaly detection, resulting in a decrease in overall VAD accuracy. In terms of efficiency, Gemini-1.5-flash emerges as the fastest model with CoT prompting, whereas VILA-13b, previously the fastest, likely loses this advantage due to difficulties in processing longer prompts.

From the comparison between two prompt frames under zero-shot and CoT prompting, we observe that a feasible way to stably enhance MLLM VAD performance is to focus on anomaly detection while enriching the prompt with contextual information about anomalies in smart home scenarios. This strategy helps compensate for the models' inherent limited understanding of anomalies.

E.2. Evaluation on Video Understanding of MLLMs

From Figure 7 and Figure 8 in the main paper, we analyze the five failure types where MLLMs failed to generate correct video description and reasoning. Additionally, we examine the distribution of MLLM outcomes for video description and reasoning across three ground-truth anomaly tags, i.e., `Normal`, `Abnormal`, and `Vague Abnormal`, as shown in Figures 23 and 24, respectively. The possible outcomes are defined as follows: (1) Correct: the MLLM's response matches the annotated description or reasoning; (2) Error: the MLLM generates "nan" or nonsensical information; (3) Incorrect: there is at least one mismatch between the MLLM output and human annotation.

For video description, over 1000 MLLM outputs are incorrect from the top three MLLMs, whereas over half of the reasoning outputs are correct. This discrepancy is likely because the description tends to include more detailed information

compared to the reasoning, as illustrated in Figure 10, making it more challenging for MLLMs to match every detail in the descriptions. The error rates for the three models follow the same ranking for both description and reasoning: Gemini-1.5-pro exhibits the highest error rate, followed by Claude-3.5-sonnet, with GPT-4o showing the least, indicating the relative stability of GPT-4o in response generation. The proportion of videos with correct descriptions across MLLMs remains consistent between `normal` and `abnormal` videos. However, the proportion of correct reasoning decreases progressively from `normal` to `abnormal` and further to `vague abnormal`. This trend highlights the limited understanding MLLMs have of smart home anomalies in our dataset, particularly for more ambiguous cases.

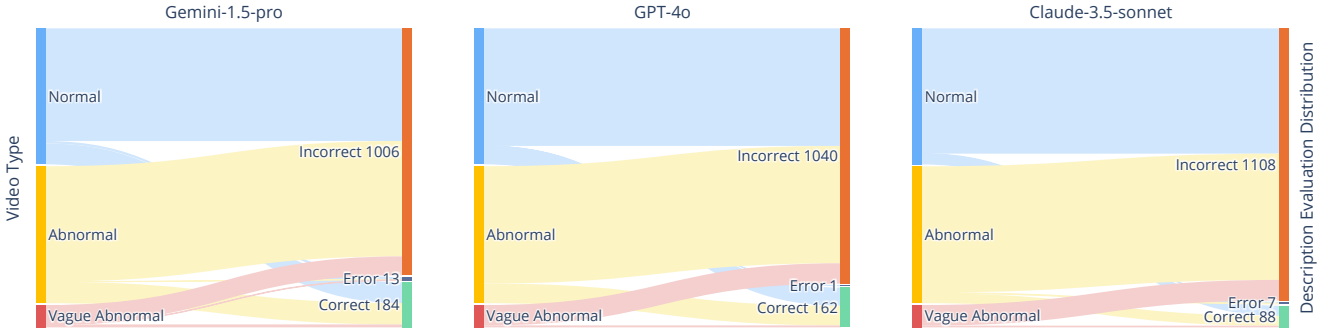

Figure 23. Distribution of video outcomes for the top three MLLMs' description compared to human-annotated description across different video anomaly tags.

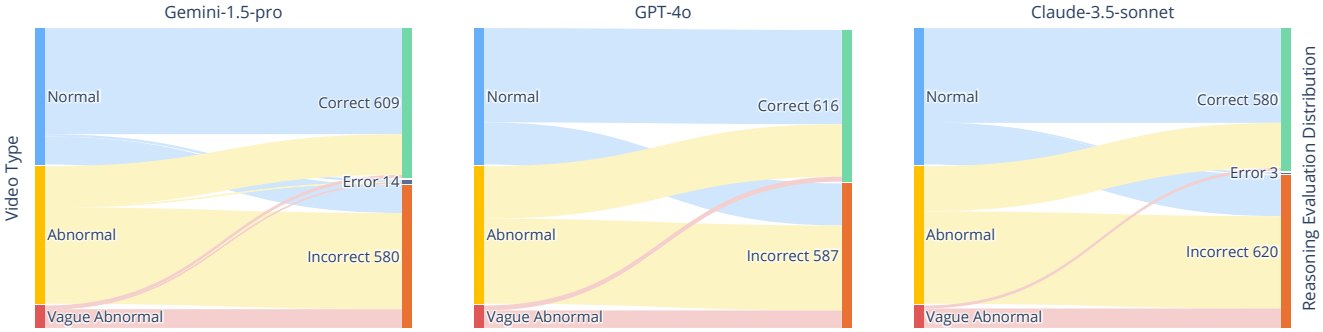

Figure 24. Distribution of video outcomes for the top three MLLMs' reasoning compared to human-annotated reasoning across different video anomaly tags.