RB-FT: Rationale-Bootstrapped Fine-Tuning for Video Classification

Anonymous ACL submission

Abstract

Vision Language Models (VLMs) are becoming increasingly integral to multimedia understanding; however, they still struggle with domain-specific video classification tasks, particularly in cases with limited data. This stems from a critical *rationale gap*, where sparse domain data is insufficient to bridge the semantic distance between complex spatio-temporal content and abstract classification labels. To bridge the gap, we propose a two-stage self-improvement paradigm without new annotations. First, we prompt the VLMs to generate detailed textual rationales for each video, compelling them to articulate the domain-specific logic. The VLM is then fine-tuned on these self-generated rationales, utilizing this intermediate supervision to align its representations with the nuances of the target domain. Second, conventional supervised fine-tuning (SFT) is performed on the task labels, achieving markedly higher effectiveness as a result of the model's pre-acquired domain reasoning. Extensive experiments on diverse datasets demonstrate that our method significantly outperforms direct SFT, validating self-generated rationale as an effective, annotation-efficient paradigm for adapting VLMs to domain-specific video analysis.

1 Introduction

Automated video analysis is increasingly vital in modern industrial settings, offering substantial gains in efficiency, safety, and quality control (Tang et al., 2025). In manufacturing, for instance, vision systems monitor assembly lines to detect subtle product defects with superhuman consistency (He et al., 2023), optimize workflows (Ravindran, 2023), and enforce safety protocols by ensuring compliance with protective equipment standards (Saurez et al., 2023). Despite their importance, these applications present a significant challenge for modern computer vision models (Shaar et al., 2025). The video data from specialized industrial environments, which are characterized by repetitive processes and minute visual anomalies, differ drastically from the general, human-centric video datasets (e.g., Kinetics (Kay et al., 2017)) used to train prominent models, creating a substantial "domain gap" that hinders out-of-the-box performance.

The current state-of-the-art in visual analysis (Luo et al., 2025; Ke et al., 2025; Hong et al., 2024; Ruthardt et al., 2024) is led by Large Vision-Language Models (LVLMs), which are pre-trained on vast internet-scale datasets of image-text pairs to learn a rich, generalizable understanding of the world (Radford et al., 2021; Li et al., 2024a; Zhou et al., 2024; Xu et al., 2024). Their ability to perform zero-shot classification by aligning visual information with natural language prompts makes them incredibly powerful (Li et al., 2025; Spinaci et al., 2025; Baia et al., 2025; Singh et al., 2024; Lavoie et al., 2024). However, this reliance on general-domain knowledge becomes a critical weakness when deploying these models in specialized industrial contexts. The visual and linguistic chasm between web data and industrial video leads to a sharp decline in performance (Abdullah et al., 2025; Rädsch et al., 2025; Fang et al., 2024; Li et al., 2024b; Parashar et al., 2024; Jiang et al., 2024). An LVLM's understanding of an abstract label like "defective" is built from diverse web images of torn clothes or dented cans. When faced with a hairline fracture on a turbine blade, the model fails not just because the visual features are unfamiliar, but because they do not align with its pre-existing, generalized concept of "defective", rendering the classification labels ineffective.

The standard method for adapting a pre-trained model to a new domain is supervised fine-tuning (SFT) with a labeled dataset (Zhang et al., 2023b). While direct SFT can provide modest improvements, its effectiveness is often limited by a deeper

issue: a semantic gap between the complex video content and the sparse classification labels. High-level labels like "pass" or "fail" provide a weak supervisory signal, making it difficult for the model to learn the intricate visual patterns of the new domain. This forces the model into an inefficient learning scenario, where it must simultaneously master the visual grammar of a new environment (e.g., unique lighting, materials, and motion) and map it to a simple label. This often leads to overfitting on superficial correlations rather than a robust, causal understanding of the task, thereby limiting performance gains (Hu et al., 2025; Niu et al., 2025; Han et al., 2024; Moenck et al., 2024).

To address these limitations, we introduce a novel rationale-bootstrapped two-stage fine-tuning framework. Our approach decomposes the difficult adaptation problem into two simpler, sequential steps. In the first stage, we employ a self-improvement mechanism inspired by Chain-of-Thought (CoT) prompting (Wei et al., 2022). We prompt the LVLM to generate detailed, step-by-step textual descriptions (or rationales) for unlabeled videos from the target domain. This creates a rich dataset of video-text pairs that we use for the initial fine-tuning stage, effectively teaching the model to "speak the language" of the target domain by grounding its visual understanding in descriptive text. This aligns with recent self-training methods where models learn from their own high-confidence, rationale-augmented outputs (Huang et al., 2022). In the second stage, once the model has been equipped with a robust, domain-specific visual representation, we perform a standard SFT using the actual classification labels. By decoupling representation learning from classification, our framework enables the model to first build a strong foundation of the new domain's visual concepts before learning the final, high-level task. Our experiments demonstrate that this method significantly enhances classification performance in specialized domains.

The main contributions are threefold:

- We first provide an in-depth analysis of the failure modes of direct SFT to domain adaptation for video classification, demonstrating how an intermediate rationale-generation stage effectively mitigates these issues.
- We propose the *Rationale-Bootstrapped Two-Stage Fine-Tuning* (RB-FT) framework, a novel fine-tuning paradigm that effectively adapts LVLMs to new domains by first bridging the semantic gap with self-generated rationales, followed by task-specific tuning.
- We conduct a comprehensive empirical study on challenging video classification benchmarks, showing that RB-FT yields significant performance gains over zero-shot, direct SFT, and other strong baselines.

2 Related Works

2.1 Evolution of Video Classification

The field of video classification underwent a significant evolution over several decades. Early methods progressed from manual descriptors, such as 3D-SIFT (Scovanner et al., 2007) and HOG3D (Klaser et al., 2008), to trajectory-based systems like Improved Dense Trajectories (iDT) (Wang and Schmid, 2013), which served as a primary baseline before the rise of deep learning. Neural networks eventually replaced manual feature design: two-stream models separated appearance from motion, 3D convolutional networks (C3D) learned joint spatial and temporal patterns (Tran et al., 2015), and hybrid models combined different network types to manage temporal sequences end-to-end (Donahue et al., 2015). Subsequent architectures addressed long-range context and processing speed, including Temporal Segment Networks (TSN), Inflated 3D ConvNets (I3D), and multi-rate designs like SlowFast (Feichtenhofer et al., 2019). These developments were further accelerated by the availability of large-scale datasets such as Kinetics (Carreira and Zisserman, 2017).

2.2 Transformer-based Models for Video Classification

Transformer architectures fundamentally advanced video understanding by using attention mechanisms to model long-range relationships in data, frequently outperforming previous convolutional methods (Bertasius et al., 2021; Arnab et al., 2021; Liu et al., 2022). In parallel, vision-language training introduced scalable supervision through natural language (Radford et al., 2021; Jia et al., 2021), which researchers successfully adapted to the video domain to allow for broader classification capabilities. More recently, the field shifted toward systems that integrated visual data with large language models to support open-ended reasoning and combined audio-visual analysis (Liu et al., 2023; Li et al., 2023; Awadalla et al., 2023; Achiam et al.,

2023; Hurst et al., 2024; Comanici et al., 2025). Specifically, methods like LLaVA-CoT (Xu et al., 2025) improved this approach by training models to perform multi-stage reasoning. Together, these advancements established a foundation for flexible video classification with high adaptability across different tasks.

2.3 Self-Improvement Under Domain Shift

While Vision Language Models (VLMs) exhibited strong generalization, they often degraded under significant distribution shifts, where abstract labels are insufficient to capture domain-specific semantics (Kay et al., 2017). Consequently, direct supervised fine-tuning (SFT) often led to overfitting or failed to bridge the semantic gap. To mitigate this, recent research pivoted toward "self-improvement" strategies that used Chain-of-Thought (CoT) prompting and model-generated explanations to provide dense supervisory signals (Huang et al., 2022; Wei et al., 2022; Wang et al., 2022). In particular, approaches such as distilling rationale from larger models (Zhang et al., 2023a) and reflective self-training (Cheng et al., 2024) demonstrated that incorporating intermediate reasoning steps significantly improves multimodal reasoning and transferability. Building on these insights, we propose Rationale-Bootstrapped Fine-Tuning (RB-FT). RB-FT adopts a two-stage paradigm in which visual representations are first aligned via detailed model-generated rationales, followed by label-based optimization. By decoupling representation learning from classification, RB-FT offers a robust alternative to direct SFT in domain-shifted contexts.

3 Method

Problem Formulation. The central problem addressed in this work is video classification under a significant domain shift. Let M_θ represent a pre-trained Vision-Language Model (VLM) with parameters θ. We denote the target domain dataset as $\mathcal{D}_{target}$, which is partitioned into a training set $\mathcal{D}_{target}^{train}$ and a testing set $\mathcal{D}_{target}^{test}$: $\mathcal{D}_{target}^{train} = \{(v_i, y_i)\}_{i=1}^{N_{train}}, \mathcal{D}_{target}^{test} = \{(v_j, y_j)\}_{j=1}^{N_{test}}$ where v represents a video instance and $y \in \{1, \ldots, C\}$ denotes the corresponding class label. The distribution of $\mathcal{D}_{target}$ is assumed to differ substantially from the data used to pretrain M_θ.

Given a sample $\mathbf{s} = (V, T)$ consisting of a video V and text T, the VLM processes each modality through specific encoders. The video input is temporally subsampled via $\mathcal{S}_\tau$ and spatially decomposed into patch tokens via $\Pi_{p,\ell}$, before being encoded by ϕ_{vid}. Concurrently, the text input is tokenized and embedded via ϕ_{text}. These multimodal tokens are augmented with positional embeddings ($\mathbf{p}$) and modality-type embeddings ($\mathbf{m}$), then concatenated into a unified sequence $\mathbf{X}$:

$$\mathbf{X} = \text{concat}\Big(\underbrace{\phi_{\text{vid}} \circ \Pi_{p,\ell} \circ \mathcal{S}_\tau(V) + \mathbf{p}^{\text{vid}} + \mathbf{m}^{\text{vid}}}_{\text{Video Embeddings}}, \underbrace{E \circ \phi_{\text{text}}(T) + \mathbf{p}^{\text{text}} + \mathbf{m}^{\text{text}}}_{\text{Text Embeddings}} \Big) \tag{1}$$

This sequence $\mathbf{X}$ is subsequently fed into the backbone Large Language Model (LLM). Our objective is to learn fine-tuned parameters θ^* using $\mathcal{D}_{target}^{train}$ that maximize classification accuracy on $\mathcal{D}_{target}^{test}$ while preserving the model's inherent multimodal reasoning capabilities.

The proposed RB-FT framework achieves this by decomposing the adaptation process into two sequential stages. The first stage focuses on domain-semantic alignment through self-generated rationales, while the second stage focuses on task-specific classification. Both stages utilize the same training cohort. The overall pipeline is shown in Figure 1.

Stage 1: Rationale-Enhanced Self-Improvement. The first stage of the RB-FT framework bridges the semantic gap between the generalist pre-trained model and the specialist target domain. This is achieved by creating and leveraging a dense, descriptive supervisory signal generated by the model itself, effectively guiding the model to learn the visual language of the new domain. This stage can be understood as a novel form of self-supervised pretext task, native to the capabilities of modern LVLMs. Traditional self-supervised learning (SSL) in video relies on pretext tasks like predicting clip order or temporal paces to force a model to learn meaningful representations without human labels. The goal of these tasks is to instill an understanding of intrinsic data properties, such as temporal coherence or motion dynamics. The rationale generation process serves a similar purpose but operates at a much higher semantic level. The pretext task becomes "generating" a coherent, detailed, and factually grounded textual description of this video." To succeed at this task, the model is compelled

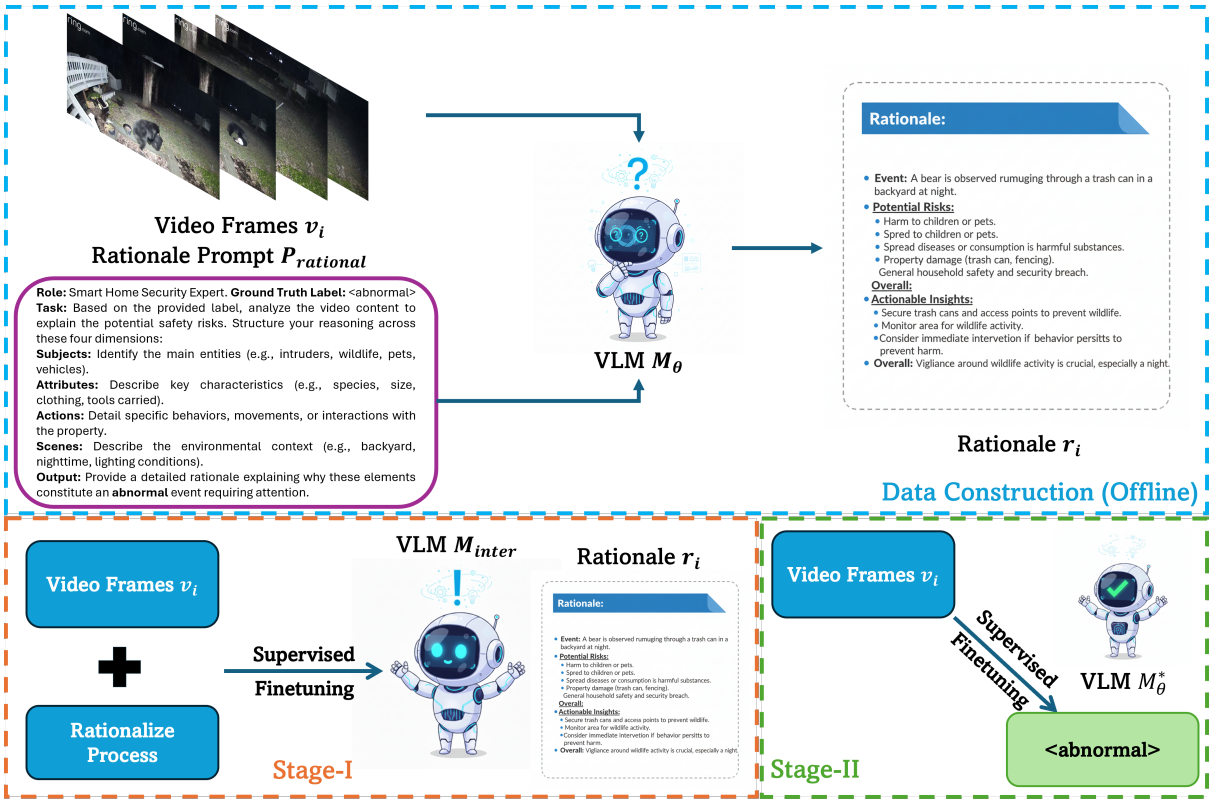


Figure 1: **Overview of the proposed Rationale-Bootstrapped Fine-Tuning Framework.** The pipeline consists of three phases: (Top) **Offline Data Construction**: A pre-trained VLM M_θ is leveraged to generate detailed textual rationales r_i for training videos using a structured prompt $P_{rationale}$. This prompt conditions the model to adopt a specific expert persona (e.g., Smart Home Security Expert) and analyze the video across four semantic dimensions: subjects, attributes, actions, and scenes. (Bottom Left) **Stage-I (Rationale-Enhanced Self-Improvement)**: The model is supervised fine-tuned to generate these domain-specific rationales, producing an intermediate model M_{inter} with enhanced reasoning capabilities. (Bottom Right) **Stage-II (Task-specific Label Alignment)**: The model undergoes a second stage of fine-tuning to predict the final ground-truth labels (e.g., `<abnormal>`), yielding the final optimized model M_θ^*.

to learn to recognize the domain-specific objects, understand their complex temporal and spatial relationships, and accurately map these visual concepts to language. Unlike traditional SSL tasks that yield a single scalar loss, this pretext task generates a rich, structured textual output that serves as a powerful and dense supervisory signal for fine-tuning.

Rationale Generation. The framework initiates by leveraging the base VLM, M_θ, to synthesize a textual rationale, r_i, for each video instance v_i within the training subset $\mathcal{D}_{target}^{train}$. This is achieved through a structured prompting strategy that bridges the domain gap. We construct a specific prompt, $P_{rationale}$, which first conditions the model to adopt the persona of a "Smart Home Security Expert," thereby aligning the generation with the specific safety-oriented requirements of the target domain.

To elicit high-quality, step-by-step reasoning, $P_{rationale}$ explicitly instructs the model to decompose its analysis across four semantic dimensions: (1) **Subjects**: identifying primary entities such as wildlife, intruders, or vehicles; (2) **Attributes**: detailing key characteristics like species, size, or clothing; (3) **Actions**: capturing dynamic behaviors, movements, and interactions with the property; and (4) **Scenes**: describing environmental contexts such as lighting conditions or backyard settings. The example is shown in Figure 1.

Formally, for each video v_i, the rationale is generated via $r_i = M_\theta(v_i, P_{rationale})$. This offline process yields an intermediate rationale-augmented dataset, $\mathcal{D}_{rationale}^{train} = \{(v_i, r_i)\}_{i=1}^{N_{train}}$, which pairs raw visual inputs with rich, self-generated semantic supervision tailored to the target distribution.

Intermediate Fine-Tuning. With the rationale dataset, $D_{rationale}^{train}$ constructed, the next step is to perform an initial supervised fine-tuning of the base

model M_θ. The objective of this intermediate SFT is to align the model's visual representations more closely with the detailed, domain-specific textual descriptions which has just been generated. The model is trained to minimize the standard auto-regressive language modeling loss over the rationales, conditioned on the corresponding videos. The loss function for this stage, $\mathcal{L}_{\text{rationale}}$, is given by:

$$\mathcal{L}_{\text{rationale}} = -\sum_{i=1}^{N_{\text{train}}} \log P(r_i \mid v_i;\, \theta)$$

M_{inter} is obtained by minimizing the $\mathcal{L}_{\text{rationale}}$.

Stage 2: Task-specific Label Alignment. The second stage of the RB-FT framework fine-tunes the domain-adapted intermediate model, M_{inter}, for the final downstream classification task. This stage uses the original ground-truth dataset, $D_{target}^{train} = \{(v_i, y_i)\}_{i=1}^{N_{train}}$.

A second round of SFT is performed, starting from the weights of M_{inter}. The objective is to minimize the standard cross-entropy loss for classification. The prompt for this stage is a simple classification query. The loss function, $\mathcal{L}_{\text{classify}}$, is:

$$\mathcal{L}_{\text{classify}} = -\sum_{i=1}^{N_{\text{train}}} \log P(y_i \mid v_i;\, \theta_{inter})$$

This fine-tuning stage is hypothesized to be significantly more effective than direct SFT on the base model M_θ. Because M_{inter} has already developed a feature space that is well adapted to the visual nuances of the target domain, the optimization landscape for the classification task is much more favorable. The model is no longer burdened with the dual challenge of learning the domain's visual language and the classification task simultaneously. Instead, the final SFT stage primarily learns to map the already meaningful and domain-aligned features to the discrete set of class labels. This separation of concerns substantially reduces the risk of catastrophic forgetting of the model's core reasoning abilities and mitigates the tendency to overfit to spurious correlations in the limited labeled data. The final output of this stage is the fully adapted model, M_θ^*.

4 Experiments

We evaluate our proposed RB-FT framework on two domain-specific video classification benchmarks: SmartHome-LLM (abnormal vs. normal daily activities) and MultiHateClip (hateful vs. normal video memes). We adopt Qwen2-VL-7B-Instruct (Wang et al., 2024b) and Qwen2.5-VL-7B-Instruct (Bai et al., 2025) as our backbone VLMs. We focus on the following three research questions, which collectively examine the central hypothesis of our paper, that bootstrapping domain-specific rationale understanding provides a strictly more transferable and more semantically stable representation prior to domain SFT:

RQ1: Quantitative Effect. Does RB-FT consistently outperform direct supervised fine-tuning (Direct-SFT) across different backbones and datasets under the same training budget?

RQ2: Factorization and Ablation. Which component(s) of RB-FT contribute the most to the final gains? In particular: (1) Where should the rationale be placed in the prompt? and (2) Does mixing self-generated rationales benefit more than partially using ground-truth rationales?

RQ3: Causal Interpretability. Does RB-FT change "what the model looks at"? For example, are attention maps qualitatively and quantitatively more causally grounded (e.g., more sensitive to masking of semantically key objects) than Direct-SFT?

4.1 Experimental Setup

Datasets. We evaluate our framework on two domain-specific video classification datasets. **SmartHome-LLM Benchmark** (Zhao et al., 2025) is a video anomaly detection benchmark featuring $1,203$ smart home clips across seven categories (e.g., Wildlife, Senior Care). Annotations include anomaly tags, detailed descriptions, and reasoning. The dataset's primary challenges lie in the subtlety of anomalies, high contextual diversity, and the requisite common-sense reasoning to distinguish normal from abnormal events. **MultiHateClip** (Wang et al., 2024a) is a multilingual benchmark for hateful video detection, comprising $2,000$ videos annotated as hateful, offensive, or normal. We utilize the English-language subset for our experiments. The benchmark's key challenges include the need for cultural nuance, the inherently multimodal nature of hate speech (encompassing visual, audio, and textual elements), and the fine-grained distinction between "hateful" and "offensive" content.

Implementation Details. We implement the RB-FT pipeline using the Qwen2-VL-7B-

Instruct (Wang et al., 2024b) and Qwen2.5-VL-7B-Instruct (Bai et al., 2025) models as our base VLMs. More details are provided in the Appendix A.

Baselines and Metrics. To establish a robust performance benchmark, we compare the proposed RB-FT framework with several baselines to rigorously validate its effectiveness. Specialized Video Models: UniFormerV2 (Li et al., 2022) and InternVideo2 (Wang et al., 2024c), which represent the state-of-the-art in models designed specifically for video understanding tasks. General-Purpose VLMs: GPT-4o-mini, GPT-4o (Hurst et al., 2024), Gemini-2.5-Flash/Pro (Comanici et al., 2025), which are powerful, proprietary large multimodal models. We also compare against the zero-shot performance of our base models and the standard Direct-SFT approach. Performance is evaluated using **Accuracy (Acc)** for overall classification performance and **F1-score** to assess the balance between precision and recall for each class, which is particularly important for datasets with imbalanced classes, such as MultiHateClip.

4.2 Main Quantitative Results

The experimental results, detailed in Table 1 and Table 2, demonstrate the consistent effectiveness of the RB-FT framework across diverse benchmarks. On the SmartHome-LLM dataset, RB-FT applied to the Qwen2.5-VL-7B backbone achieves a state-of-the-art accuracy of 82.65%. This represents a significant 6.63 percentage-point increase over standard fine-tuning and a 27.55 percentage-point improvement over the zero-shot baseline. A key highlight is the F1 (Normal) metric, which rose from 51.55% to 76.06%, indicating that our rationale-bootstrapped alignment successfully bridges the semantic gap. By training the model to first generate descriptive logic for video content, the framework cultivates a robust understanding of standard patterns, allowing our open-source model to mitigate the performance gap and even surpass high-resource proprietary models like Gemini-2.5-Pro.

This efficacy extends to the MultiHateClip dataset, where RB-FT reaches 71.00% accuracy and doubles the F1-score for the minority "Hateful" class. These results show that the dense supervisory signals from self-generated rationales help the model identify distinct features for underrepresented categories that are typically overlooked during standard training. Furthermore, the consistent improvements across both Qwen2-VL and Qwen2.5-VL architectures confirm that RB-FT is a versatile solution for adapting to new domains. By providing a structured reasoning path, the framework addresses core challenges in video understanding that persist across different model architectures.

4.3 In-Depth Analysis and Discussion

We conduct a series of ablation studies to understand the mechanisms behind RB-FT's effectiveness. These studies collectively reveal a coherent narrative about how the framework operates: the specific composition of the rationale supervision enables a scalable self-generation process, which results in a more interpretable and robust final model. All the ablation studies are conducted on the SmartHome-LLM dataset.

Ablation Study: Impact of Rationale Composition. We examine how the composition of the rationales affects the learning. The term rational is fixed and refers only to the generated reasoning description. We compare three supervision formats that differ only in whether and where the final classification appears relative to the rationale, while keeping the rationale's content unchanged. We denote P as the input prompt, R as the rational, and C as the final class label. The formats are P+R with no explicit class label, P+C+R with the class label placed before the rational, and P+R+C with the class label placed after the rational. During fine-tuning, the model is trained to produce the target in the specified order, and at evaluation, we elicit outputs in the same order.

Across our experiments, the P+R setting achieves the best overall performance, surpassing both variants with an explicit class label. The gains arise because supervision that focuses on rationales directs learning toward the reasoning steps rather than reproducing label tokens. Omitting the label discourages shortcutting, reduces label leakage, and encourages the model to ground its predictions in event descriptions, object interactions, and causal relations. Placing the label at the beginning promotes post hoc justification, whereas putting it at the end still provides a strong cue that can be memorized. Training that focuses solely on the rationales yields a denser, more informative signal and improves generalization, calibration, and robustness for tasks that require multi-step reasoning.

Ablation Study: Ratio of Self-Generated Ra-

Table 1: Quantitative results on the SmartHome-LLM benchmark. Models are categorized into proprietary (closed-source) and open-source systems. For each metric, the best performance across proprietary models is underlined, while the best result within the open-source category is **bolded**.

Category	Model	Method	Accuracy (%) ↑	F1 (Normal) ↑	F1 (Abnormal) ↑
Proprietary (Closed-source)	GPT-4.1	Zero-shot	70.53	69.23	71.72
	Gemini-2.5-Flash	Zero-shot	73.33	70.79	75.47
	GPT-4o	Zero-shot	62.76	63.32	62.18
	Gemini-2.5-Pro	Zero-shot	73.47	70.11	76.15
Open-source (Open-weight)	UniFormerV2	Supervised	70.12	38.02	78.95
	InternVideo2	Supervised	68.70	35.40	79.10
	Qwen2-VL-7B	Zero-shot	57.65	59.11	56.08
		Direct-SFT	69.39	36.17	79.87
		RB-FT (Ours)	80.61	74.67	84.30
	Qwen2.5-VL-7B	Zero-shot	55.10	60.36	48.24
		Direct-SFT	76.02	51.55	84.07
		RB-FT (Ours)	**82.65**	**76.06**	**86.40**

Table 2: Quantitative results on the MultiHateClip benchmark. Models are categorized into proprietary (closed-source) and open-source systems. For each metric, the best performance across proprietary models is underlined, while the best result within the open-source category is **bolded**.

Category	Model	Method	Acc. (%) ↑	F1 (Norm) ↑	F1 (Hate) ↑	F1 (Offen) ↑
Proprietary (Closed-source)	GPT-4.1	Zero-shot	70.46	83.58	60.38	23.53
	Gemini-2.5-Flash	Zero-shot	70.94	83.98	18.18	40.00
	GPT-4o	Zero-shot	72.19	84.16	55.45	12.50
	Gemini-2.5-Pro	Zero-shot	75.81	84.71	61.11	0.00
Open-source (Open-weight)	UniFormerV2	Supervised	59.17	72.10	8.33	35.42
	InternVideo2	Supervised	62.72	74.85	10.91	38.06
	Qwen2-VL-7B	Zero-shot	53.25	65.33	0.00	39.37
		Direct-SFT	65.68	78.81	4.76	29.63
		RB-FT (Ours)	68.64	81.27	14.29	42.65
	Qwen2.5-VL-7B	Zero-shot	54.48	66.90	2.70	40.12
		Direct-SFT	66.86	79.92	11.11	41.35
		RB-FT (Ours)	**71.00**	**83.47**	**23.53**	**49.78**

tionales in Stage-I. We next evaluate the impact of using self-generated rationales versus human-annotated ones by training models with varying ratios of synthetic data in Stage-I. Table 4 shows a clear and positive trend: performance improves as the proportion of self-generated rationales increases, with the 100% self-generated setting yielding the best results.

This result strongly validates the premise that the model's own descriptions, even if imperfect, are a powerful and effective source of supervision. It suggests that for this domain adaptation task, the breadth of descriptive coverage across the entire training set (achieved with 100% self-generated data) is more valuable than the potentially higher quality of a small subset of human annotations. This has profound implications for annotation efficiency, proving that the rationale-based learning signal is robust enough to be self-generated at scale. The goal is not to model the distribution of rationales but to use them as a rich signal to learn better visual representations.

Ablation Study: Impact of Key Objects. To investigate whether RB-FT learns a more causal and interpretable model, we perform an object masking ablation. We identify the three most salient objects using Gemini-2.5-pro in each test video and compare the model's performance on the original frames, frames with these objects masked, and frames with random patches of equivalent area masked.

The results are shown in Table 5. For the RB-FT model, masking critical objects (Obj. Masked) causes a dramatic drop in accuracy. This drop is far more significant than that caused by masking random patches (Rand. Masked, 75.51%) and substantially larger than the drop observed for the Direct-SFT model (69.39% to 67.86%). This provides strong evidence that the RB-FT model has learned to ground its classifications in semantically meaningful objects. These very objects would be explicitly named in the generated rationales. In contrast, the Direct-SFT model relies more on diffuse, superficial correlations across the entire frame, making it less sensitive to the removal of specific objects. This demonstrates that the rationale-based super-

Table 3: Ablation Study on the Composition of the Rationales.

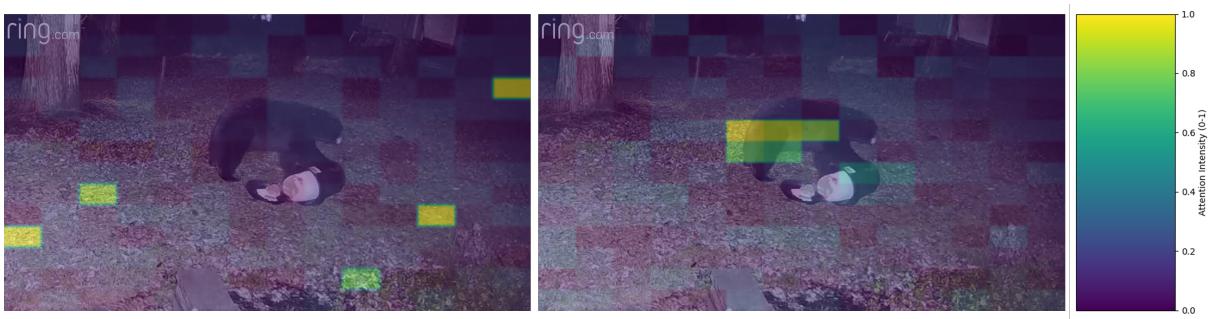
Model	Rationale	Acc (%) ↑	F1 (Normal) ↑	F1 (Abnormal) ↑
Qwen2-VL-7B	P+C+R	79.59	73.33	83.47
	P+R+C	80.61	73.97	84.55
	P+R	80.61	74.67	84.30
Qwen2.5-VL-7B	P+C+R	81.10	74.90	85.30
	P+R+C	81.65	75.10	85.70
	P+R	**82.65**	**76.06**	**86.40**

Table 4: Ablation Study on the Ratio of Self-Generated Rationales in Stage-I.

Model	Ratio	Acc (%) ↑	F1 (Normal) ↑	F1 (Abnormal) ↑
Qwen2-VL-7B	20%	71.43	53.20	81.90
	60%	74.49	56.80	82.70
	100%	**80.61**	**74.67**	**84.30**
Qwen2.5-VL-7B	20%	76.53	59.80	85.40
	60%	78.57	62.10	85.90
	100%	**82.65**	**76.06**	**86.40**

Table 5: Ablation Study on the Key Objects.

Model	Paradigm	Input	Acc. (%) ↑	F1 (Normal) ↑	F1 (Abnormal) ↑
Qwen2-VL-7B	Direct_SFT	Rand. Masked	66.84	21.69	78.96
		Obj. Masked	67.86	18.18	80.00
		Ori. Frames	69.39	36.17	79.87
	RB-FT	Rand. Masked	75.51	63.08	81.68
		Obj. Masked	66.33	40.00	76.60
		Ori. Frames	**80.61**	**74.67**	**84.30**



(a) Attention Map of Direct-SFT (b) Attention Map of RB-FT

Figure 2: Comparative visualization of attention maps between the baseline Direct-SFT model (a) and our proposed RB-FT model (b). The heatmaps (purple to yellow) represent attention intensity, scaled from 0 to 1. The RB-FT model (b) demonstrates significantly improved focal accuracy, concentrating high-intensity attention on the salient subjects (the black bear and the trash can), whereas the Direct-SFT model (a) exhibits diffuse attention, failing to localize the critical regions of interest.

vision in Stage 1 produces a model that is more accurate, more interpretable, and causally aligned with the task.

4.4 Attention Mode Analysis

To qualitatively assess the impact of our proposed RB-FT, we visualized the attention distributions of the baseline Direct-SFT model and our RB-FT model on a representative sample, shown in Figure 2. The resulting attention maps illustrate the models' ability to localize key objects within the input. As evidenced in the comparison, the Direct-SFT approach (a) produces a scattered and unfocused attention pattern, erroneously assigning importance to irrelevant background elements while failing to identify the critical subjects. In stark contrast, our proposed RB-FT method (b) demonstrates exceptional effectiveness, generating a highly concentrated attention map that precisely localizes the key entities in the scene. This visualization confirms that the RB-FT technique successfully guides the model's focus to semantically critical regions, suppressing background noise and enabling a more robust, accurate understanding of the scene's content.

5 Conclusion

This work introduces Rationale-Bootstrapped Fine-Tuning (RB-FT) to address the semantic gaps inherent in specialized video domains where standard supervised fine-tuning fails. By decoupling domain-semantic alignment from task classification via self-generated rationales, RB-FT provides dense intermediate supervision that aligns visual representations with domain-specific concepts. Extensive experiments demonstrate that this approach significantly outperforms standard baselines, producing models that are both robust and interpretable.

6 Limitations

While RB-FT effectively addresses the semantic gap in domain-specific video classification by decoupling representation alignment from task tuning, it still imposes certain constraints. First, the framework's performance is intrinsically linked to the underlying Large Vision-Language Model's (LVLM) capacity for structured reasoning; in highly specialized domains where the base model lacks foundational visual concepts, the initial bootstrapped signal may require more intensive prompting to ensure high-fidelity supervision. Second, the two-stage fine-tuning process, while central to our performance gains, introduces a more complex training profile and longer computational time than single-stage direct supervised fine-tuning (Direct-SFT). Finally, while our evaluation demonstrates robust results on established benchmarks like SmartHome-LLM and MultiHateClip, the scalability of self-generated rationales for long-video understanding or multi-modal inputs involving synchronized audio remains an area for future investigation.

7 Ethics Statement

The deployment of automated video analysis systems, particularly in sensitive contexts such as home security and content moderation, necessitates careful ethical consideration. RB-FT promotes transparency by requiring the model to articulate its decision-making logic through textual rationales, thereby providing a human-interpretable audit trail for its classifications. We utilize publicly available research benchmarks and advocate for the application of this framework within strict privacy-preserving protocols that comply with local data protection regulations. Furthermore, by grounding the model's focus on specific subjects, attributes, and actions, our approach aims to mitigate the risk of the model relying on superficial or biased correlations inherent in pre-trained foundation models. Ultimately, we emphasize that this framework is intended to serve as a supportive tool for human-in-the-loop decision-making in safety-critical industrial and social environments.

References

Raiyaan Abdullah, Yogesh Singh Rawat, and Shruti Vyas. 2025. isafetybench: A video-language benchmark for safety in industrial environment. In *ICCV*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *ICCV*.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, and 1 others. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Alina Elena Baia, Alessio Xompero, and Andrea Cavallaro. 2025. Zero-shot image privacy classification with vision-language models. *arXiv preprint arXiv:2510.09253*.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *ICML*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.

Kanzhi Cheng, Yantao Li, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. 2024. Vision-language models can self-improve reasoning via reflection. *arXiv preprint arXiv:2411.00855*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. In *NeurIPS*.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *ICCV*.

Jinwei Han, Zhiwen Lin, Zhongyisun Sun, Yingguo Gao, Ke Yan, Shouhong Ding, Yuan Gao, and Gui-Song Xia. 2024. Anchor-based robust finetuning of vision-language models. In *CVPR*.

Jianben He, Xingbo Wang, Kam Kwai Wong, Xijie Huang, Changjian Chen, Zixin Chen, Fengjie Wang, Min Zhu, and Huamin Qu. 2023. Videopro: A visual analytics approach for interactive video programming. *TVCG*.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, and 1 others. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Yangliu Hu, Zikai Song, Na Feng, Yawei Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2025. Sf2t: Self-supervised fragment finetuning of video-llms for fine-grained understanding. In *CVPR*.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

Yao Jiang, Xinyu Yan, Ge-Peng Ji, Keren Fu, Meijun Sun, Huan Xiong, Deng-Ping Fan, and Fahad Shahbaz Khan. 2024. Effectiveness assessment of recent large vision-language models. *Visual Intelligence*.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, and 1 others. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Fucai Ke, Joy Hsu, Zhixi Cai, Zixian Ma, Xin Zheng, Xindi Wu, Sukai Huang, Weiqing Wang, Pari Delir Haghighi, Gholamreza Haffari, and 1 others. 2025. Explain before you answer: A survey on compositional visual reasoning. *arXiv preprint arXiv:2508.17298*.

Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. 2008. A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British machine vision conference*. British Machine Vision Association.

Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mahmoud Assran, Andrew Gordon Wilson, Aaron Courville, and Nicolas Ballas. 2024. Modeling caption diversity in contrastive vision-language pretraining. *arXiv preprint arXiv:2405.00740*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. 2022. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024a. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*.

Xinhao Li, Zhenpeng Huang, Jing Wang, Kunchang Li, and Limin Wang. 2024b. Videoeval: Comprehensive benchmark suite for low-cost evaluation of video foundation model. *arXiv preprint arXiv:2407.06491*.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *CVPR*.

Ziyang Luo, Haoning Wu, Dongxu Li, Jing Ma, Mohan Kankanhalli, and Junnan Li. 2025. Videoautoarena: An automated arena for evaluating large multimodal models in video analysis through user simulation. In *CVPR*.

Keno Moenck, Duc Trung Thieu, Julian Koch, and Thorsten Schüppstuhl. 2024. Industrial language-image dataset (ilid): adapting vision foundation models for industrial settings. *Procedia CIRP*.

Ke Niu, Zhuofan Chen, Haiyang Yu, Yuwen Chen, Teng Fu, Mengyang Zhao, Bin Li, and Xiangyang Xue. 2025. Creft-cad: Boosting orthographic projection reasoning for cad via reinforcement fine-tuning. *arXiv preprint arXiv:2506.00568*.

Shubham Parashar, Zhiqiu Lin, Tian Liu, Xiangjue Dong, Yanan Li, Deva Ramanan, James Caverlee, and Shu Kong. 2024. The neglected tails in vision-language models. In *CVPR*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Tim Rädsch, Leon Mayer, Simon Pavicic, A Emre Kavur, Marcel Knopp, Barış Öztürk, Klaus Maier-Hein, Paul F Jaeger, Fabian Isensee, Annika Reinke,

and 1 others. 2025. Bridging vision language model (vlm) evaluation gaps with a framework for scalable and cost-effective benchmark generation. *arXiv preprint arXiv:2502.15563*.

Arun A Ravindran. 2023. Internet-of-things edge computing systems for streaming video analytics: Trails behind and the paths ahead. *IoT*.

Jona Ruthardt, Gertjan J Burghouts, Serge Belongie, and Yuki M Asano. 2024. Better language models exhibit higher visual alignment. *arXiv preprint arXiv:2410.07173*.

Enrique Saurez, Harshit Gupta, Henriette Roger, Sukanya Bhowmik, Umakishore Ramachandran, and Kurt Rothermel. 2023. Utility-aware load shedding for real-time video analytics at the edge. *arXiv preprint arXiv:2307.02409*.

Paul Scovanner, Saad Ali, and Mubarak Shah. 2007. A 3-dimensional sift descriptor and its application to action recognition. In *ACM MM*.

Eitan Shaar, Ariel Shaulov, Gal Chechik, and Lior Wolf. 2025. Adapting to the unknown: Training-free audio-visual event perception with dynamic thresholds. In *CVPR*.

Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. 2024. Learn" no" to say" yes" better: Improving vision-language models via negations. *arXiv preprint arXiv:2403.20312*.

Gianmarco Spinaci, Lukas Klic, and Giovanni Colavizza. 2025. Benchmarking vision-language and multimodal large language models in zero-shot and few-shot scenarios: A study on christian iconography. *arXiv preprint arXiv:2509.18839*.

Hengzhu Tang, Zefeng Zhang, Zhiping Li, Zhenyu Zhang, Xing Wu, Li Gao, Suqi Cheng, and Dawei Yin. 2025. Multi-branch collaborative learning network for video quality assessment in industrial video search. *arXiv preprint arXiv:2502.05924*.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*.

Han Wang, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee. 2024a. Multihateclip: A multilingual benchmark dataset for hateful video detection on youtube and bilibili. In *ACM MM*.

Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *ICCV*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, and 1 others. 2024c. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step. In *ICCV*.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2024. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *TPAMI*.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and 1 others. 2023b. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Xinyi Zhao, Congjing Zhang, Pei Guo, Wei Li, Lin Chen, Chaoyue Zhao, and Shuai Huang. 2025. Smarthome-bench: A comprehensive benchmark for video anomaly detection in smart homes using multi-modal large language models. In *CVPR*.

Kaiwen Zhou, Kwonjoon Lee, Teruhisa Misu, and Xin Wang. 2024. Vicor: Bridging visual understanding and commonsense reasoning with large language models. In *Findings of the ACL*.

A Implementation Details

The videos are sampled at 1 FPS with a maximum resolution of 360×420 pixels. For both fine-tuning stages, we use a cosine learning rate scheduler with a 3% warm-up rate and a weight decay of 0.1. The learning rates are set to 1×10^{-5} for the language model and merger components and 2×10^{-6} for the vision tower. We enable gradient checkpointing and clip gradients at a norm of 1.0. Training is performed with a global batch size of 16. Both Stage-I and Stage-II are trained for a single epoch to mitigate the risk of overfitting on the specialized datasets. All experiments are conducted on 8 NVIDIA H100 GPUs.

B Additional Experimental Results

To rigorously evaluate the causal necessity of semantic relevance during the rationale-bootstrapped phase, we conduct a correlation study on the SmartHome-LLM benchmark using two redesigned adversarial prompting strategies: P_{less} (Peripheral Context) and P_{least} (Atmospheric Narrative).

In the P_{less} scenario, the model acts as a **Contextual Property Observer**, generating descriptions focused on the general surroundings and architectural layout. This strategy captures the spatial context of the video but deliberately omits the specific human-centric actions or potential risks required for anomaly detection. In the P_{least} scenario, the model adopts the persona of an **Atmospheric Journalist**, focusing exclusively on sensory and stylistic elements such as lighting transitions, weather effects seen through windows, and camera stability. While this rationale is "related" to the video file itself, it provides near-zero semantic utility for distinguishing between normal and abnormal behavioral patterns.

Using LLM-as-a-judge (specifically a GPT-4o evaluator (Hurst et al., 2024)), we quantified the alignment between these rationales and the classification topic, defined as the causal mapping between visual evidence and ground-truth behavioral labels. As detailed in table 6, the task-correlation score dropped from 0.88 for our original rationale to 0.42 for P_{less} and 0.15 for P_{least}. The results reveal a strictly monotonic decline in performance as the semantic correlation decreases. Notably, the P_{least} setting performs slightly below Direct-SFT (75.82% vs 76.02%), suggesting that fine-tuning on semantically irrelevant descriptions can introduce noise that distracts the model from task-specific visual features. In contrast, the substantial 6.63 percentage-point gain of RB-FT is fundamentally driven by high-correlation causal reasoning. This confirms that the framework's efficacy stems from bridging the "rationale gap" through task-aligned semantic supervision rather than simple data exposure.

Table 6: Correlation Analysis: Impact of Rationale Relevance on SmartHome-LLM Performance using Qwen2.5-VL-7B-Instruct. Task correlation scores are determined via LLM-as-a-judge on a scale of 0 to 1.

Setting	Task Corr. ↑	Acc. (%) ↑	F1 (Norm) ↑	F1 (Abnorm) ↑
RB-FT (Original)	**0.88**	**82.65**	**76.06**	**86.40**
P_{less} (Static Scene)	0.42	78.42	64.15	84.92
P_{least} (Metadata)	0.12	76.95	54.30	84.18
Direct-SFT	0.00	76.02	51.55	84.07

C Claim of Use of AI Assistants

We use AI assistants to help us polish the writing.