The Not-So-Secret Life of Dogs

Bea Lim
Stanford University
Stanford, CA
bealim@stanford.edu

Abstract

This project presents a system for fine-grained pet behavior monitoring using video captioning and summarization. I fine-tune SmolVLM2-2.2B, a lightweight vision-language model, on 91 labeled motion-triggered home clips to improve captioning of dog activities. The small size enables efficient fine-tuning and supports potential real-world deployment on consumer-grade devices. Compared to the baseline, METEOR and BERTScore F1 scores improve from 0.541 to 0.564 and 0.753 to 0.785, respectively. Captions from daily recordings are aggregated into behavior summaries using prompt-based inference. While summaries are often informative, occasional hallucinations reflect limitations in the LLM component of SmolVLM2, which can generate actions not present in the input. This work demonstrates a novel application of video-language models for passive home monitoring, and lays groundwork for future extensions in real-time, behavior-aware pet care systems.

1. Introduction

1.1. Motivation

Pet monitoring systems today often fail to bridge the gap between motion detection and semantic understanding. Like many pet owners, I've returned home to chewed slippers or missing socks, with only a fragmented pile of motion-triggered videos to explain what happened. These cameras—while useful—overwhelm users with unstructured clips and lack interpretability. A summarized activity report would offer a practical, digestible update on pet behavior, especially during long absences.

This project explores how vision-language models (VLMs) can be applied to generate structured behavioral summaries from raw home surveillance footage, transforming hours of passive video into a compact and meaningful activity report. The motivation stems from the desire to convert low-level sensing into high-level insight—making pet care more proactive, personalized, and efficient.

1.2. Problem Statement

The core problem is the lack of semantic interpretation in consumer-grade home monitoring. While devices like Ring or Blink provide motion-triggered video, they offer no description of what occurred, where, or whether it was noteworthy. This results in a disconnect between the abundance of visual data and actionable behavioral insight. To address this, I propose a two-stage pipeline that performs:

Input: Motion-triggered home video clips (30–60 seconds), each showing one or more dog behaviors.

Model: A fine-tuned SmolVLM2-2.2B vision-language model.

Output:

1. Natural language captions describing the dog's behavior per video.
2. A structured daily summary synthesized from the full set of captions.

The system performs two primary tasks:

1. [Video Captioning] Generate fine-grained, interpretable descriptions of dog behavior from surveillance clips. This is responsible for generating accurate and factual event records.
2. [Summarization] Aggregate these captions into a human-readable log, highlighting major events and potential concerns. This is responsible for delivering a practical and meaningful report to pet owners.

While the summarization component involves prompting the LLM portion of SmolVLM2, I include it here for completeness. Improving summarization quality is considered out of scope for this project and is left to future work focused on language grounding and LLM behavior tuning.

To maintain clarity, I annotate subsections of this report with [Video Captioning] and [Summarization] tags where appropriate.

2. Related Works

Pet activity monitoring has evolved from sensor-based and video classification systems to recent advances in multimodal vision-language models (VLMs). This work extends the field by using a fine-tuned VLM to generate open-ended captions and structured summaries of dog behavior from real-world home video. Below, I review key prior efforts in commercial systems, sensor-based approaches, and vision-based or multimodal classifiers.

2.1. Commercial Systems

Furbo Dog Nanny [9] and PetCube [10] offers a proprietary behavior alert system for pet owners, generating daily summaries of barking, movement spikes, or visitor activity. However, this functionality is tightly coupled to Furbo's and Petcube's hardware ecosystem and limited to a small set of hardcoded events. This project seeks to enable similar behavior summarization using an open-source VLM and commodity hardware, with richer, caption-level interpretation.

Tomofun [4], the company behind Furbo, provides a blog post describing the cloud backend architecture needed to support large-scale video ingestion and real-time alert delivery. While it does not describe ML techniques for behavior recognition, the infrastructure design reflects growing commercial interest in scalable pet monitoring solutions.

2.2. Sensor-Based and IoT Monitoring

[3] Hsieh et al. (2022) present a lightweight pet monitoring system using human activity recognition techniques over heterogeneous sensor networks. While their focus is on using sensor fusion rather than visual understanding, their work highlights the growing importance of low-power, real-time inference for in-home pet monitoring.

Chen et al. [5] introduced an IoT-based interactive system that relayed alerts and enabled user interaction, but used basic rule-based pipelines.

These sensor-heavy pipelines highlight the historical focus on structured inputs and rule-based alerts, which lack the flexibility and generalization required for nuanced behavior understanding.

2.3. Vision-Based/Multimodal Behavior Classification

Other approaches frame pet monitoring as a vision-based classification task. Kim and Moon [7] applied CNNs to image classification of pet behaviors like lying, sitting, and walking using webcam input. While their system achieved high accuracy for static behavior detection, it could not model temporal sequences or generate descriptive language.

A. Lin et al. [11] has explored multi-modal behavior understanding using dog-mounted egocentric cameras and audio to classify specific reactions to environmental stimuli (e.g., Sit, Stand, Walk, Smell), this approach focuses on home surveillance from the human perspective—using stationary, consumer-grade cameras placed in typical living spaces. Rather than classifying discrete reactions to stimuli, my system generates fine-grained, natural-language descriptions and daily summaries of a pet's activities throughout the entire home, targeting interpretability and actionable insights for everyday pet owners. This addresses the gap between highly controlled, sensor-rich egocentric setups and practical, scalable pet monitoring solutions deployable in real-world home environments.

Recent work by Martin et al. [12] uses advanced computer vision to quantify tail-wagging as an emotional marker in controlled settings. In contrast, my approach leverages consumer-grade cameras to capture diverse dog behaviors throughout the home and generates natural-language summaries for practical, real-world use by pet owners.

While prior work (e.g., Atif et al. [8]) focuses on controlled, single-room monitoring with specialized action recognition models and detailed visualization tools for expert analysis, my approach targets real-world, multi-area home environments using lightweight, consumer-ready vision-language models. By placing cameras in typical household locations and generating owner-focused natural language summaries, my system bridges the gap between technical behavior monitoring and interpretable, actionable reports for everyday pet owners, addressing usability and deployment challenges not covered by previous systems.

3. Dataset

This dataset centers around Amaru, a one-year-old Shiba Inu who serves as the sole subject of this study. Amaru's natural behaviors, captured in a typical home environment, provide a rich foundation for enabling the investigation of fine-grained dog behavior recognition.

Figure 1 Example frames of Amaru, illustrating the diversity of locations, poses and lighting conditions captured in the home environment.

3.1. Data Collection [Video Captioning]

The dataset consists primarily of short video clips (captured via Blink and Ring home monitoring systems) and still images (taken with iPhone) of Amaru in various parts of the house. To realistically simulate typical pet owner behavior and maximize generalizability, cameras were placed in easy, non-permanent locations such as shelves, tables, or other elevated surfaces, to reflect how most owners would monitor their pets without drilling or complex installation.

The primary filming locations included the home's most frequented areas, such as the living room and the hallway leading to the front door, ensuring coverage of spaces where Amaru spends the majority of his time and encounters a range of daily situations.

Clips were handpicked to capture a diverse array of activities (e.g., chewing, playing, sleeping, sitting, zoomies), interactions with various objects (e.g., toys, bones, socks, shoes), and environmental contexts (e.g., at the door, by a window, on the couch, near valuable or potentially hazardous items like laptops and computers).
Each sample was manually annotated with structured ground truth labels:

- Activity (chewing, sleeping, etc.)
- Object of interaction (sock, toy, etc.), if any
- Location (on the couch, near the window, at the door, etc.)

The dataset is split into two: fine-tuning (training) dataset and testing dataset. Please refer to Table 1 for the distribution of activity categories in the two datasets.

3.1.1 Testing Dataset:

A curation of small, representative dataset of 30 samples. This dataset serves as the gold-standard benchmark for evaluating the performance of baseline and fine-tuned models.

3.1.2 Training Dataset:

91 labelled video examples with a variety of activity classes. These will be used to fine-tune the pretrained VLM used in this project.

Table 1 Distribution of activity categories in the video captioning datasets

Activity Category	Examples	# Train Samples	# Test Samples
Play	Chewing toys/bone	10	25
Rest	Sleeping/Lying on floor/couch/bed	12	30
Waiting	Sitting/Lying down in front of the door	4	1
Eat/Drink	Drinking water	1	3
Zoomies	Running energetically	1	5
Wandering	Walking around the room	1	7
Mischief	Flailing stolen sock in air	1	8
Possible Discomfort	Panting, sneeze attack	0	2
Foraging	Clawing under the couch to retrieve something	0	3
Self-grooming	Licking paws	0	6
Total		30	91

3.2. Data Collection [Summarization]

To assess the summarization task, I collected motion-triggered video clips from my Ring home cameras capturing Amaru during a single real-world session home alone for 5 hours. Each video clip is named with its start timestamp. This dataset is intentionally limited, as it serves as an initial case study for evaluating the summarization module, which depends solely on the LLM component of SmolVLM2 and is considered out-of-scope for extensive evaluation in this work.

4. Technical Approach

SmolVLM2-2.2B, a lightweight vision-language model designed for efficient video understanding, was selected for this project. The choice was motivated by two key factors: deployment feasibility and fine-tuning flexibility.
SmolVLM2-2.2B is significantly smaller than most state-of-the-art VLMs, making it well-suited for potential deployment on home monitoring systems (e.g., Blink or Ring cameras), where inference speed and limited compute are critical constraints. Its compact size enables faster iteration and better suitability for real-time or near-real-time use, which is important for building a practical pet activity summarization system.

The high-level approach to develop this system was as follows:

1. Assess baseline (pre-trained) SmolVLM2-2.2B video captioning performance on the test dataset → Section 4.1.3
2. Fine-tune SmolVLM2-2.2B to improve captioning using the training dataset → Section 4.1.4
3. Evaluate the fine-tuned model's captioning performance on the test dataset
4. Run the fine-tuned model on the summarization dataset (a single 5-hour home-alone session) to obtain a list of timestamped video captions
5. Summarize list of captions into a short report using the fine-tuned SmolVLM2

4.1. Technical Approach [Video Captioning]

The objective of the video captioning was focused on factual accuracy of the scenes from the inputted videos. Behavior is not inferred here yet as the summarization task will handle this when it based on the full list of activities carried out by the dog.

4.1.1 Objective and Prompt Strategy

The objective of the video captioning is to produce concise, factually accurate descriptions of Amaru's observable actions, context, and interactions in each video. To guide the model, I used the following explicit prompt:

> "What is Amaru (the dog) doing in this video, and where is he doing it? Describe only observable actions, his body language or expression if visible, and any relevant object or setting he interacts with. Be concise and factual. Do not describe the dog's appearance, breed, or color."

This prompt was designed to focus the model on owner-relevant behavioral details, avoiding redundant information (such as breed or color or that he's a dog) to ensure that the generated captions remained practical.

4.1.2 Evaluation

Performance of both the baseline (pre-trained) and fine-tuned SmolVLM2-2.2B models were evaluated on the original 30 sample test dataset using the standard metrics outlined below.

BERT Score

This metric captures semantic similarity using contextual embeddings, making it tolerant to synonyms and phrasing variation (ideal for evaluating fine-grained behavior descriptions).

$$R_{\text{BERT}} = \frac{1}{|x|}\sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \ , \quad P_{\text{BERT}} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \ , \quad F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

Figure 2 BERT Score Equations

- Precision (P): How much of the generated caption is semantically relevant to the ground truth.
- Recall (R): How much of the ground truth caption is captured by the generated caption.
- F1: Harmonic mean of precision and recall.

METEOR Score

This metric emphasizes exact and stemmed word overlap with synonym matching via WordNet. It is useful for evaluating if the generated captions cover the key content words and details present in the ground truth labels.

$$Fmean = \frac{10PR}{R+9P} \ , \quad Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams\_matched}\right)^3 \ , \quad Score = Fmean * (1 - Penalty)$$

Figure 3 METEOR Score Equations

Using both these metrics together provides a well-rounded assessment of how faithfully the model's captions reflect the intended information in the ground truth.

4.1.3 Baseline Performance

Before fine-tuning, the pre-trained SmolVLM2-2.2B model was evaluated on the 30-example test dataset of home videos of Amaru (described earlier in Section 3.1.1). While the model demonstrated strong generalist capabilities, its performance was moderate (see Table 3 for quantitative results and for comparison against the fine-tuned model). Qualitative review revealed some shortcomings in its ability to describe pet-specific behaviors in a detailed and practical way, such as specifying exactly what the dog was doing, what object it was interacting with, and where (per the prompt in Section 4.1.1). Some examples of these notable shortcomings:

- Captions sometimes over-described the scene (see Appendix A for a sample)
- Object of interaction were inconsistently captured (See Appendix A for a sample)
- Some outputs included hallucinated scene elements These limitations highlighted the need for domain-specific adaptation.

4.1.4 Fine-tuning Strategy

The SmolVLM2-2.2B vision-language model was fine-tuned on the 91 labeled home videos of Amaru (described earlier in Section 3.1.2), using HuggingFace Transformers and PyTorch. This section details the experimental setup, hyperparameter choices, and rationale, as well as resource considerations that shaped the workflow.

4.1.4.1 Resource Constraints and Considerations

Fine-tuning multimodal models like SmolVLM2-2.2B can still be relatively memory-intensive. While training on an NVIDIA A100 (40GB VRAM), I frequently encountered CUDA out-of-memory (OOM) errors. Diagnostic output indicated that PyTorch reserved substantial memory that was not always efficiently allocated for active computation, likely due to fragmentation and the dynamic memory allocation patterns of video batch processing (memory usage fluctuates a lot as with variable-length video batches).

To address these issues:

- Batch Size was chosen to be 2, with gradient accumulation enabled (with steps=2) to simulate a larger effective batch size while minimizing peak memory usage. This allowed for stable training to be maintained without sacrificing effective throughput.
- The paged_adamw_8bit optimizer (from bitsandbytes) was used, which is specifically designed to minimize memory footprint when training large models, by sharding optimizer states and supporting 8-bit quantization of weights and gradients.

Although parameter-efficient techniques (such as LoRA, or QLoRA which updates only a small subset of parameters, enabling training without modifying the full base model) exist, I opted to fine-tune the full SmolVLM2-2.2B model for several reasons:

- Domain Gap: While the baseline model could generally describe the scene, it often produced generic or redundant details (e.g., "Amaru is a brown dog…") that were less actionable for pet owners. In contrast, the ground truth labels focused on providing specific, owner-relevant information such as the dog's current

activity and its implications for welfare, highlighting the need for fine-tuning to produce more practically useful, behavior-focused captions.
- Resource Availability: With access to a high-memory GPU (A100 40GB), I was able to attempt full-model fine-tuning.

4.1.4.2 Hyperparameter and Training Configuration

In addition to the configurations mentioned in the previous section (to be more memory-efficient), the following configurations were selected:

Table 2 Fine-Tuning Hyperparameters and Configurations

Hyperparameters / Training Config	Rationale
Epochs = 1	Due to the small training dataset size of 91, only one epoch was used for training to avoid overfitting.
Learning Rate = 1e-4 Weight Decay = 0.01	Standard learning rates for large vision-language models, with weight decay to mitigate overfitting
Optimizer = AdamW	Allows for both custom weight decay (for better generalization), and custom adaptive learning rates (for faster and more stable convergence).

Additionally, a custom data collator constructed batches by applying a chat-style prompt template to combine video content and ground-truth captions. It handled padding for both text and video inputs per batch, ensuring efficient batch processing.

4.2. Technical Approach [Summarization]

The summarization system generates structured high-level activity reports using the fine-tuned SmolVLM2 model. Rather than relying on a separate, general-purpose LLM (such as GPT-4, Llama 3, or Qwen-LLM) for summarization, this approach prompts the same fine-tuned SmolVLM2 directly to condense its own list of captions into a high-level log. This unified design was intended to simplify deployment and to ensure tight integration between captioning and summarization.

A list of timestamped captions is re-fed into the fine-tuned SmolVLM2 model with an instruction-oriented prompt, enabling the model to output a concise, prioritized summary of the dog's activities.

4.2.1 Objective and Prompt Strategy

Each motion-triggered clip is first processed by SmolVLM2 to generate a caption and paired with its timestamp. These timestamped captions are then grouped and re-fed into SmolVLM2 with an instruction prompt that guides the model to:

1. Merge repeated or continuous behaviors
2. Emphasize priority events (e.g., mischief, distress)
3. Minimize redundancy

The exact summarization prompt given to the model:

"From the list of timestamped activity captions of my dog, Amaru, please summarize my dog's day. I do not need it to be fine-grained but just high-level activities which can be grouped, but:
(1) if there was any mischievous activities (stealing socks, chewing household objects), this needs to be reported each time. do not falsely report this.
(2) If there is any distress (pacing, excessive barking, limping, vomiting, barking), this needs to be reported.
(3) If he is hanging around the door, it can be regarded as 'waiting' for us.
(4) If he looks directly at the camera, it can be regarded as a 'he took a selfie' and can be reported.
The output should only be 2-4 short sentences. Mention what activity he did more of."

This approach ensures that the summary highlights critical incidents while also providing a high-level, interpretable view of the dog's day for pet owners.

4.2.2 Evaluation

Given the limited scope of this project (focusing mainly the on the vision model component), evaluation in this study was primarily qualitative. The generated summary was manually reviewed for completeness, clarity, and practical value to a pet owner. Special attention was paid to whether high-priority events (such as mischief or distress behaviors) identified in the ground truth captions were faithfully included in the summary, reflecting real-world owner concerns. This approach was chosen due to the dataset's small size and the proof-of-concept focus of this module.

Extensive quantitative evaluation, including broader human surveys or automated scoring, is left to future work as the summarization component is not the primary focus of this study.

5. Results

5.1. Results [Video Captioning]

5.1.1 Fine-Tuning

The fine-tuning of SmolVLM2 for video captioning was conducted for a single epoch (as previously stated in Table 2), resulting in 22 training steps (derived from 91 training examples with an effective batch size of 4, using a batch size of 2 and gradient accumulation over 2 steps).

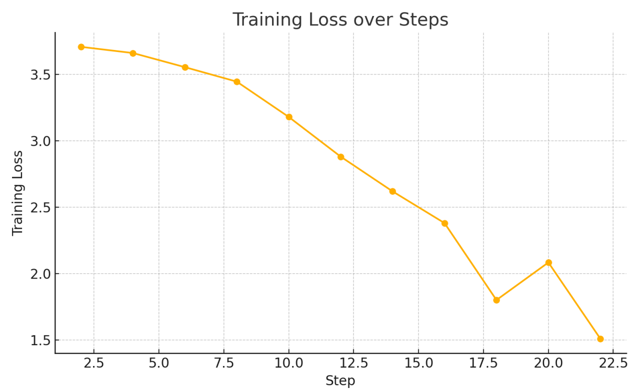


Figure 4 Training loss over steps during fine-tuning of SmolVLM2 for video captioning, showing a smooth decrease as the model learns from the training data

As shown in the plot, training loss was logged every 2 steps. The loss decreased smoothly from 3.71 to 1.51 over the course of training, demonstrating stable optimization and indicating that the model was successfully learning from the training data. The small bump observed near step 21 is expected and primarily due to two factors:

- Small batch size: Each batch may contain quite different examples, resulting in higher fluctuation (noise) in the loss values from step to step.
- Small dataset: With fewer samples overall, random variation in batch composition has a larger effect on training loss.

5.1.2 Captioning

The captioning performance of the SmolVLM2-2.2B model before and after fine-tuning was evaluated on the test dataset. Quantitative results are shown in *Table 3*, using the metrics previously described in Section 4.1.2.

Table 3 Video Captioning Performance Changes

Metric	**Baseline**	**After Fine-tune**	**Delta**	**Interpretation**
METEOR	0.5412	0.5641	+4.2%	Moderate gain implying word choice improvement by the model.
BERTScore Precision	0.7542	0.8172	+8.4%	Decent gain suggesting that the model generates more semantically accurate captions.
BERTScore Recall	0.76	0.7618	+0.2%	Unchanged indicating that the model captures about the same coverage of ground truth content.
BERTScore F1	0.7535	0.7851	+4.2%	Moderate gain, due to increased precision.

With only 91 supervised examples, these results are quite encouraging. The fine-tuned model became noticeably more precise in its descriptions while maintaining recall, suggesting it learned to generate more concise and behaviorally relevant captions.

Upon manual review of the generated captions, a few qualitative improvements were observed, along with a few unchanged errors are found (Appendix A presents both) like minor object misidentifications (e.g., sock vs shoe). Nonetheless, the model generally produced behaviorally accurate and contextually relevant captions.

5.2. Results [Summarization]

For the summarization task, the fine-tuned SmolVLM2 model was prompted with a list of timestamped activity captions (see Appendix B for the full list) and asked to generate a concise summary. Manual review of the generated summaries revealed several key observations:

- Prompt Sensitivity: The model's output was highly sensitive to the phrasing of the instruction prompt; even minor wording changes produced noticeably different summaries.
- Hallucination: The model frequently included details not present in the input captions, such as reporting "mischief" or "selfies with the owner" that did not actually occur.
- Generalization: In some cases, the model was able to group and summarize dominant patterns (e.g., "Amaru spent most of the day sleeping, playing, and walking around"), but still introduced inaccuracies due to overgeneralization or invented events.

Representative prompt-response pairs are provided below, with the full prompt texts and complete list of activity captions included in Appendix B.

Test Prompt 1 (see Appendix B for exact prompt) produced the following summary:

"Amaru spent most of the day sleeping, playing with toys, and exploring the living room. He also had a few moments of mischief, such as stealing socks and chewing on household objects. However, he did not show any signs of distress or overheating. Amaru took a selfie with his owner and seemed happy to be spending time with them."

Test Prompt 2 (see Appendix B for exact prompt) produced:

"Amaru spent most of the day sleeping, playing, and walking around. He took a selfie and was waiting for his owners."

Both summaries were generated from the same list of timestamped activity captions.

These issues are likely due to the fact that SmolVLM2 was only fine-tuned for factual captioning, not for robust instruction-following or grounded summarization, and likely lacks robustness to prompt phrasing, as it has not seen diverse summarization-style instructions during fine-tuning. Hallucination and prompt sensitivity are expected under such circumstances. Addressing these limitations would require further instruction tuning of the model's language head. These directions are left to future work and are further discussed in Section 6.2, as improvements to the LLM component are out-of-scope for this paper.

6. Conclusion and Future Work

6.1. Conclusion

This project demonstrates the feasibility of using a lightweight vision-language model, SmolVLM2-2.2B, for fine-grained dog behavior monitoring from consumer-grade home surveillance data. By fine-tuning on a small, carefully curated dataset, I showed improvements to the factual accuracy and owner-relevance of video captions compared to the pre-trained baseline. These results suggest that compact VLMs can be adapted for practical, real-time deployment on devices with limited compute, bridging the gap between raw video data and interpretable behavioral insights for pet owners.

Although the summarization component showed promise in generating high-level daily activity reports, it exhibited sensitivity to prompt phrasing and a tendency to hallucinate details. These issues highlight the limitations of using a captioning-tuned LLM head for robust instruction-following and grounded summarization. Addressing these challenges (through instruction tuning, and larger datasets) remains an important direction for future work.

Overall, this study provides a proof of concept for end-to-end, vision-language-driven pet monitoring and lays a foundation for further research toward dog behavior-aware, automated home care systems.

6.2. Future Work

The motivation for future work stems not only from the academic value of this project, but also from its practical utility. Being able to meaningfully interpret Amaru's daily activities when I'm away has proven both insightful and useful in my own home.

Although the scope this quarter was limited by time, several promising directions for improvement and broader utility are proposed below:

6.2.1 Dataset Improvements (Section 3)

- Expand and Balance the Dataset: Collect additional video clips to increase the total (both test and train) dataset size and ensure a more balanced distribution across all activity categories in both training and testing splits.
- Data Augmentation: Explore augmentation techniques to improve model robustness given the limited original data (black and white video frames which is what is collected from Blink and Ring cameras when there is low light)
- Summarization Scenarios: Gather more real-world, multi-session data for the summarization component to better reflect diverse daily routines, and to enable thorough performance evaluation.

6.2.2 Fine-Tuning Enhancements (Section 4.1.4)

- Targeted Masking: Refine the fine-tuning process by masking all tokens except the ground-truth caption/answer during training. This will help focus the learning signal, prevent overfitting to prompt phrasing, and make maximal use of limited labeled data.

6.2.3 Summarization Evaluation and Improvement (Section 4.2.2, 5.2)

- Human Evaluation: Conduct user studies or surveys to collect human ratings on summary faithfulness, clarity, and usefulness, providing a more rigorous assessment of real-world value.
- Broader Benchmarking: Expand evaluation to more home-alone scenarios for a robust, generalizable assessment of summarization quality.
- Instruction Tuning: Further train the language head of SmolVLM2 for instruction-following and factual summarization, possibly using parameter-efficient techniques like LoRA or QLoRA to preserve captioning ability and avoid catastrophic forgetting.
- Constrained Decoding: Implement control tokens or pattern-based constraints to ensure that summaries only reference activities present in the input list, reducing hallucination.

Pursuing these improvements would enable a more reliable, generalizable, and user-friendly system for automated pet behavior monitoring and summarization.

7. Acknowledgements

I would like to thank Amaru, a very good boy, for being the star and unwitting participant of this project. His daily adventures inspired this work and made the research process genuinely enjoyable. I also thank my course instructors and teaching assistants for their feedback and support throughout the quarter.

References

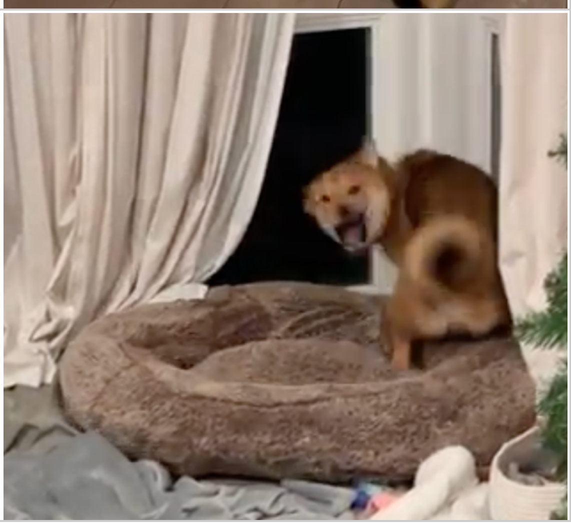
[1] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating Text Generation with BERT. In International Conference on Learning Representations (ICLR), 2020. https://arxiv.org/pdf/1904.09675

[2] S. Banerjee and A. Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005. https://www.cs.cmu.edu/~alavie/METEOR/pdf/Banerjee-Lavie-2005-METEOR.pdf

[3] Y.-H. Hsieh, S.-W. Tsai, Y.-Y. Liu, and S.-F. Su. Lightweight Pet Monitoring with Human Activity Recognition using Heterogeneous Sensor Networks. In 2022 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), pages 1–2, 2022. https://ieeexplore-ieee-org.stanford.idm.oclc.org/stamp/stamp.jsp?tp=&arnumber=9774398
[4] Tomofun. Backend Cloud Architecture: A Pet Monitoring System's Scalable Backend Design. Tomofun Tech Blog, 2020. https://medium.com/tomofun-tech-blog/backend-cloud-architecture-1dbfe3da971c
[5] C.-Y. Chen, K.-T. Song, and J.-H. Chen. Cloud Pet Care: A Pet Behavior Monitoring and Interactive System Based on IoT. In 2014 IEEE International Conference on Consumer Electronics (ICCE), pages 165–166, 2014. https://ieeexplore-ieee-org.stanford.idm.oclc.org/stamp/stamp.jsp?tp=&arnumber=6977451
[6] C. Nguyen, A. Nguyen, and C. Hillar. SmolVLM: A Surprisingly Small Vision-Language Model for Zero-Shot Video Understanding. arXiv preprint arXiv:2504.05299, 2024. https://arxiv.org/pdf/2504.05299
[7] J. Kim and N. Moon. "Dog Behavior Recognition Based on Multimodal Data from a Camera and Wearable Device", Applied Sciences, vol. 12, no. 6, 3199, 2022. https://www.mdpi.com/2076-3417/12/6/3199
[8] O. Atif, J. Lee, D. Park, and Y. Chun. "Behavior-Based Video Summarization System for Dog Health and Welfare Monitoring" Sensors, vol. 23, no. 6, p. 2892, 2023. https://www.mdpi.com/1424-8220/23/6/2892
[9] Furbo Nanny, Available at https://furbo.com/us/pages/furbo-nanny?srsltid=AfmBOoqvkeNescJFAfN8rOzuSQX8-bbthz7BTwPHerufKWb9usJzWSef
[10] PetCube, Available at https://petcube.com/play/
[11] A. Lin and M. Eastburn. "Understanding Dog Behavior Through Visual and Auditory Sensing Using Machine Learning." Journal of Student Research, vol. 12, no. 4, 2023. https://www.jsr.org/hs/index.php/path/article/view/5801/2677
[12] M. Martin, et al. "Automated Depth Sensing-Based Computer Vision for Dog Tail Wagging Interpretation." Proceedings of ACI 2024, Glasgow, United Kingdom, December 02–05, 2024.

Appendix

Appendix A

Thumbnail	(Ground Truth) Caption	(Baseline) Caption	(Finetuned) Caption
	Amaru is sitting in front of the door.	Amaru is a brown dog sitting in front of a white door. He is wearing a collar and looking at the door.	Amaru is sitting in front of a white door.
	Amaru is on his dog bed, running energetically in circles trying to catch its tail.	Amaru is sitting on a brown dog bed in a living room. He is chewing on a toy and then moves to a different spot on the bed.	Amaru is lying on a dog bed, chewing on a toy.

Legend:

	OK Caption
	Partially OK Caption
	NOK Caption

Appendix B

Table I Detailed Summarization Result: Inputs and outputs for the summarization task are shown below. Differences between the two prompts are highlighted in blue.

Input List of Timestamped Captions for Summarization

13-52-10	Amaru is walking on the floor.
13-00-50	Amaru is sitting on the floor in a living room.
13-17-04	Amaru is lying on the floor.
13-14-42	Amaru is lying on the floor next to a bed.
13-03-35	Amaru is lying on the floor in front of a bed.
13-37-22	Amaru is lying on the floor next to a bed.
13-48-26	Amaru is lying on the floor next to a bed.
13-00-07	Amaru is walking around the living room.
17-01-33	Amaru is lying on a dog bed in the living room.
13-35-29	Amaru is lying on the floor next to a bed.
12-59-27	Amaru is playing with a red ball.
13-12-47	Amaru is lying on the floor next to a bed.
12-58-22	Amaru is lying on the floor in the living room.
13-52-11	Amaru is lying on the floor.
13-02-55	Amaru is playing with a toy.
13-19-17	Amaru is lying on the floor next to a bed.
15-47-33	Amaru is lying on a dog bed in the living room.
13-31-30	Amaru is sleeping on the floor.
13-49-49	Amaru is lying on the floor next to a bed.
13-02-51	Amaru is sitting on the floor in the living room.
13-14-13	Amaru is lying on the floor next to a bed.
13-34-59	Amaru is sleeping on the floor.
12-57-00	Amaru is walking around the living room.
13-01-48	Amaru is sitting on the floor in a living room.
12-57-26	Amaru is walking on the floor.
12-59-09	Amaru is playing with a toy.
13-15-06	Amaru is lying on the floor in front of a bed.
13-17-51	Amaru is lying on the floor next to a bed.

Summarization Prompt 1	Summarization Prompt 2
From the list of timestamped activity captions of my dog, Amaru, please summarize my dog's day. I do not need it to be fine-grained but just high-level activities which can be grouped, **but with the following exceptions**: (1) if there were any mischievous activities (stealing socks, chewing household objects), this needs to be reported each time. Do not falsely report this. (2) If there is any distress (pacing, excessive barking, limping, vomiting), **note panting as a potential sign of overheating or stress, if barking at the door, mention the possibility of external stimuli (mailman);** (3) **Active Behavior (summarize patterns): Group together high-energy movement (e.g., zoomies, running, jumping); Mention toy play, exploration, or wandering (wandering may indicate curiosity, boredom, or restlessness).** (4) If he looks directly at the camera, it can be regarded as a selfie and can be summarized in the report as well. The output should only be 2-4 short sentences.	From the list of timestamped activity captions of my dog, Amaru, please summarize my dog's day. I do not need it to be fine-grained but just high-level activities which can be grouped, **but**: (1) if there were any mischievous activities (stealing socks, chewing household objects), this needs to be reported each time. Do not falsely report this. (2) If there is any distress (pacing, excessive barking, limping, vomiting, barking), **this needs to be reported.** (3) **If he is hanging around the door, it can be regarded as 'waiting' for us.** (4) If he looks directly at the camera, it can be regarded as a 'he took a selfie' and can be reported. The output should only be 2-4 short sentences. **Mention what activity he did more of.**

Output Summary 1	Output Summary 2
Amaru spent most of the day sleeping, playing with toys, and exploring the living room. He also had a few moments of mischief, such as stealing socks and chewing on household objects. However, he did not show any signs of distress or overheating. Amaru took a selfie with his owner and seemed happy to be spending time with them.	Amaru spent most of the day sleeping, playing, and walking around. He took a selfie and was waiting for his owners.