Pawgaze: A Benchmark for Fine-Grained Multimodal Analysis of Canine Behavior

Khush Attarde
Research and Development Department
Ogmen Robotics
Billerica, MA 01821
`khush.attarde@ogmenrobotics.com`

Divye Bhutani
Research and Development Department
Ogmen Robotics
Billerica, MA 01821
`divye@ogmenrobotics.com`

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated remarkable progress in zero-shot understanding of diverse inputs such as video, audio, and text. **But can they accurately understand complex animal behavior?** This challenge stems from the lack of comprehensive datasets that capture real-world animal behaviors, combining visual and auditory cues with insights into physical conditions and emotional states. To address this gap, we present **Pawgaze**, a novel benchmark for fine-grained analysis of dog activities, comprising **7,120 question–answer pairs across 923 videos**. The benchmark includes real-world dog videos paired with synchronized audio–visual, five-way multiple-choice questions requiring frame-level reasoning, interpretation of behavioral cues, and understanding of human–dog interactions. We introduce a scalable, LLM-based automated question-answer generation pipeline that is facilitated by domain expert-driven insights developed in collaboration with *canine behavior experts*. MLLM benchmarking is conducted using various proprietary MLLMS models. Experimental results and analyses indicate that closed-source MLLMs demonstrating superior zero-shot performance in multimodal understanding of canine-centered behaviors but rely heavily on prior knowledge. A detailed failure analysis highlights the challenges and opportunities for improvement. Pawgaze paves the way for extending VLM capabilities beyond traditional scene understanding tasks, with promising applications in pet-care robotics, animal health, and behavior modeling. We provide a link to the anonymized dataset here.

1 Introduction

In modern times, as the bond between humans and animals grows stronger, understanding animal behavior has gained increasing importance. Among pets, dogs are one of the most preferred companions and often require greater care. Yet, analyzing canine emotions and behaviors remains challenging due to their subtle, context-dependent, and multimodal signals. While recent Video-Language Models (VLMs) have shown strong performance in human activity recognition by leveraging visual and audio inputs [1, 2, 3, 4, 5, 6], their application to non-human animals is limited by the lack of specialized datasets.

Animal-focused datasets, such as MammAlps [7], LoTE-Animal [8], Animal Kingdom [9], CBVD-5 [10], AnimalWeb [11], KABR [12], CamoVid60K [13], CVB [14], and MammalNet [15], support behavior recognition in wildlife or livestock but largely exclude emotional states, long-term dynamics, and social interactions. Canine-specific datasets such as Stanford Dogs [16], Tsinghua Dog [17], synthetic pose recognition [18], and egocentric videos [19] primarily focus on classification or pose tasks, while the DEBIw dataset [20] offers image-based dog emotion recognition but lacks

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Non-Human Animal Communications.

temporal and multimodal depth. Overall, fine-grained audio-visual benchmarks for canine behavior understanding remain absent or unexplored; refer to Appendix A for detailed information.

Modeling progress further highlights this gap. Breed classification relies on visual CNNs and SVMs [21], yet behavior understanding demands multimodal fusion of posture, movement, and vocalizations [22, 23, 24, 25]. Systems like AmadeusGPT [26] and MouseGPT [27] showcase advances in combining pose estimation, segmentation, and open-vocabulary behavior annotations, but remain species-specific and not tailored to dogs.

To address these challenges, we introduce **Pawgaze**, a benchmark of real-world dog videos paired with synchronized audio-visual, five-way multiple-choice questions that demand temporal reasoning, behavioral cue interpretation, and human–dog interaction understanding. This dataset fills a critical gap by enabling multimodal AI research in canine communication, with promising applications in pet-care robotics, animal health, and behavior modeling.

2 Methodology

Video Collection: Videos were collected through two approaches: (1) **query-based search** using the YouTube API v3 [1] with behavior-related keywords, and (2) **curated seed selection** of predefined video IDs from open datasets or manual reviewer input. From these seeds, the pipeline expands by retrieving additional videos from the same channels and YouTube recommendations. Metadata (title, description, ID, duration, timestamp, channel ID) is stored alongside each video, downloaded via `yt-dlp` [28]. **Filtering:** Collected videos undergo automatic filtering using `Gemini-2.0-Flash`, which evaluates both metadata and visual–audio input to classify relevance (`YES`/`NO`). In early testing (765 videos), downsampling at 1 FPS with audio preserved was attempted, but later subsampling at fixed FPS without audio synchronization proved more effective for capturing filtering. Human validation of accepted/rejected samples guided prompt refinement and confirmed that the subsampled-video approach improved accuracy. The final pipeline integrates these refinements, ensuring only relevant dog activities remain while non-relevant content and metadata are discarded. Details of prompts, verification results, and configurations are in Appendix B.

Table 1: Our proposed task categories with question prototypes.

Canine Descriptive Foundations	
Behavior Profiling	*Describe the pacing behavior of the dog in the kitchen when the food in the bowl is visible.*
Posture Analysis	*Describe the dog's ear and head position when the stranger enters the park. What does this suggest about its alertness?*
Steps of Action	*Trace the steps the dog takes from noticing the toy to engaging in play with the human.*
Canine State and Purpose Understanding	
Emotion Analysis	*What emotion is the dog likely experiencing while a stranger approaches the front door?*
Contextual Interpretation	*How does leash restriction alter the dog's behavior when an unfamiliar dog enters the park?*
Causal Inference	*What event immediately triggers the dog to nudge its owner repeatedly?*
Canine Social and Relational Dynamics	
Social Interaction Analysis	*What social behavioral cues suggest that the dog is seeking comfort from the human after the loud noise?*
Comparative Behavioral Analysis	*Compare the dog's gait while walking on the sidewalk with its gait while circling in the backyard. What does this difference suggest about its emotional state?*
Interactive Loop Analysis	*Trace the loop starting when one dog growls at another near the food bowl. How does this affect the second dog's response and the first dog's stance?*

Taxonomy and Task Suites: In collaboration with canine behavior experts, we developed a taxonomy mapping observable cues (behavioral, social, postural) to over 70 behaviors, 14 emotions, and fine-grained states, structured with guidance from prior work [29] and canine behavior domain experts; details and taxonomy are provided in Appendix C. This taxonomy underpins a knowledge base that integrates object–behavior affordances, spatial ontology, postures, gaits, and vocalizations,

[1] `https://developers.google.com/youtube/v3/docs`

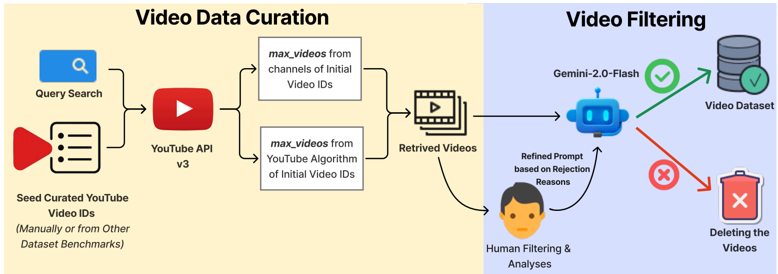


Figure 1: Video curation and filtering pipeline leveraging a Human-in-the-Loop scalable refinement process for ensuring high-quality and contextually accurate behavioral data.

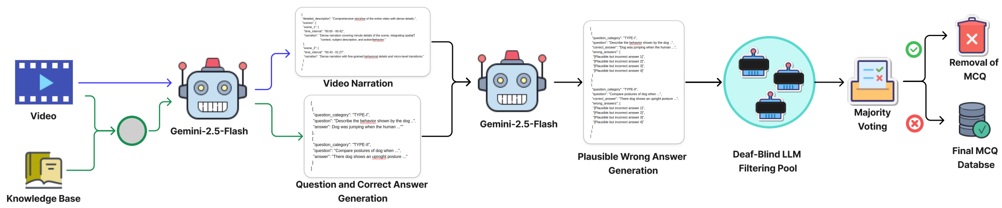


Figure 2: Multiple-choice question generation pipeline.

ensuring QA generation remains expert-grounded and systematic, details provided in Appendix C. Building on this foundation, our benchmark organizes tasks into three categories: *Canine Descriptive Foundations* (observable behaviors, postures, actions), *Canine State and Purpose Understanding* (internal states, motivations, contextual/causal factors), and *Canine Social and Relational Dynamics* (communication, interactions, relational adaptations). Representative task prototypes are summarized in Table 1.

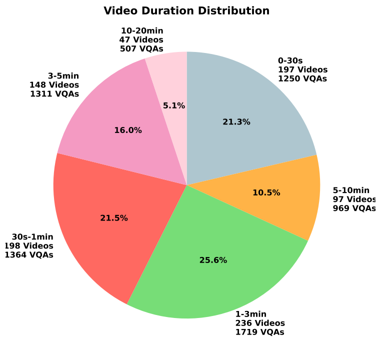
Question–Answer Dataset Generation Pipeline: **Video Narration:** Using `Gemini-2.5-Flash`, each video was processed with audio–visual inputs to produce temporally segmented narrations with timestamps. These narrations captured activity, interaction, contextual meaning and later guided distractor generation. **Question–Answer Pair Generation:** QA pairs were created with reference to the knowledge base and predefined question types (Table 1). `Gemini-2.5-Flash` generated initial questions and correct answers from videos, followed by four distractors derived from narrations, answers, and metadata (without reusing full videos). This ensured distractors required fine-grained reasoning. The final multiple-choice sets were organized per category. See Appendix D.1 and D.2 for prompts used in video narration and MCQ generation.

Deaf-Blind LLM Filtering: To exclude items solvable by prior LLM knowledge, we conduct text-only evaluations (without input videos frames) with three open-source LLMs to filter generated QA pairs via majority voting. Overall, **44.36%** of samples were discarded, with additional outliers removed by video-length duration constraints to maintain a balanced distribution in the dataset. Full prompt and configuration details and deaf-blind llm filtering performance are in Appendix D.3.

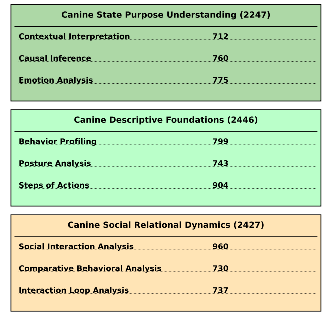
3 Benchmark Analyses and Evaluations

Pawgaze is a canine behavior-specific dataset for analyzing fine-grained behaviors, including social, emotional, and contextual cues with their interpretations. From an initial pool of **1929** videos, **923** were retained after filtering, yielding **7120** five-way multiple-choice (MCQ) pairs. The dataset spans diverse video lengths (see Figure 3) and was developed with guidance from canine behavior experts.

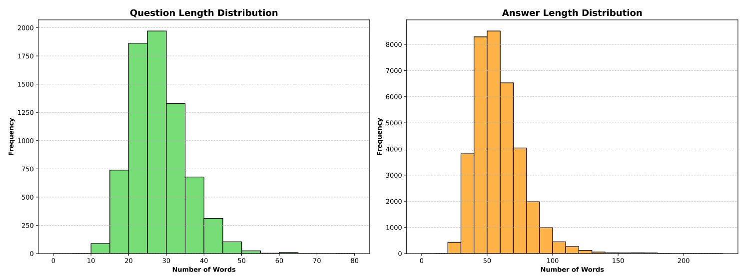
Gemini-2.0-flash Evaluations. We evaluate the Pawgaze benchmark using state-of-the-art closed source models, including `Gemini-2.0-Flash` and `GPT-4o`, their quantitative results can be seen in Table 2. Gemini-2.0-Flash is accessed via URL requests [30] and configuration selected based on initial analyses (see Appendix F.1). The results (see Appendix F.2) show higher accuracy in some categories, while others such as Steps of Actions, Emotion Analyses, and Posture Analysis remain



(a) Video lengths



(b) Canine category composition



(c) QA lengths

Figure 3: Statistical overview of the Pawgaze dataset: (a) video lengths, (b) category composition, and (c) QA lengths.

Table 2: Overall and Category-wise accuracy (in %) of closed-source models across the Pawgaze benchmark.

Model	**Frame**	Behavior Profiling	Posture Analysis	Steps of Actions	Emotion Analyses	Contextual Interpretation	Causal Inference	Social Interaction Analyses	Comparative Behavioral Analyses	Interaction Loop Analyses	**Overall**
Gemini-2.0-Flash	all	63.58	57.20	59.73	55.87	60.39	69.61	58.23	56.58	69.34	61.07
GPT-4o	32	60.00	47.90	45.80	50.86	57.96	57.58	57.24	61.89	67.95	56.07

challenging, especially in medium to longer videos. Qualitative analyses further reveal that models that rely on video frames only often construct incomplete narratives from limited frames, leading to misinterpretation. In contrast, multimodal models leverage synchronized audio and narration to provide critical contextual cues, resulting in more accurate alignment with ground truth (refer to Appendix E). Incorrect choices by Gemini-2.0-Flash stem from cue misinterpretation, visual errors, overgeneralization (see Appendix F.3 for details).

GPT-4o Evaluations. As shown in Table 2, GPT-4o achieves lower overall accuracy (56.07%) than Gemini-2.0-Flash (61.07%), with notably poor performance in Steps of Actions (45.8%) and Posture Analysis (47.9%). While GPT-4o often selects the correct answer, its chain-of-thought reasoning relies on general knowledge and option-based extrapolation rather than true frame-level understanding (Appendix G, Examples 1 and 3). In contrast, Gemini-2.0-Flash produces stepwise, frame-grounded reasoning, accurately interpreting visual cues such as hand signals, verbal commands, and detailed interactions, highlighting the importance of multimodal models for canine understanding.

4 Conclusion

Pawgaze establishes the first fine-grained benchmark for multimodal understanding of canine behavior, spanning 923 real-world videos and 7,120 QA pairs. Analyses show that while both GPT-4o and Gemini-2.0-Flash achieve competitive results, Gemini-2.0-Flash consistently outperforms GPT-4o, particularly through frame-grounded reasoning that leverages both visual and contextual cues. In contrast, GPT-4o, despite sometimes selecting correct answers, often relies on general knowledge and option-based extrapolation rather than true frame-level interpretation. Failure cases across emotion analyses, posture recognition, and stepwise actions highlight the limitations of current MLLMs in complex animal understanding. We will explore benchmarking and finetuning open-source MLLMs [31] to improve the fine-grained dog behavior understanding beyond proppreitary models. Overall, Pawgaze provides a rigorous evaluation framework, revealing both the progress and challenges in extending multimodal models to animal-centered domains, with applications in pet-care robotics, health monitoring, and behavioral modeling.

References

[1] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

[2] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023.

[3] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[4] Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark, 2024.

[5] Arsha Nagrani, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu, Nilpa Jha, Austin Myers, Xingyi Zhou, Boqing Gong, Cordelia Schmid, et al. Neptune: The long orbit to benchmarking long video understanding, 2024.

[6] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark, 2016.

[7] Valentin Gabeff, Haozhe Qi, Brendan Flaherty, Gencer Sumbul, Alexander Mathis, and Devis Tuia. Mammalps: A multi-view video behavior monitoring dataset of wild mammals in the swiss alps. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13854–13864, 2025.

[8] Dan Liu, Jin Hou, Shaoli Huang, Jing Liu, Yuxin He, Bochuan Zheng, Jifeng Ning, and Jingdong Zhang. Lote-animal: A long time-span dataset for endangered animal behavior understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20064–20075, 2023.

[9] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19023–19034, 2022.

[10] Kuo Li, Daoerji Fan, Huijuan Wu, and Aruna Zhao. A new dataset for video-based cow behavior recognition. *Scientific Reports*, 14(1):18702, 2024.

[11] Muhammad Haris Khan, John McDonagh, Salman Khan, Muhammad Shahabuddin, Aditya Arora, Fahad Shahbaz Khan, Ling Shao, and Georgios Tzimiropoulos. Animalweb: A large-scale hierarchical dataset of annotated animal faces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6939–6948, 2020.

[12] M Kholiavchenko, J Kline, M Ramirez, S Stevens, A Sheets, R Babu, N Banerji, E Campolongo, M Thompson, N Van Tiel, et al. Kabr: In-situ dataset for kenyan animal behavior recognition from drone videos. in 2024 ieee/cvf winter conference on applications of computer vision workshops (wacvw)(pp. 31e40), 2024.

[13] Tuan-Anh Vu, Ziqiang Zheng, Chengyang Song, Qing Guo, Ivor Tsang, and Sai-Kit Yeung. Camovid60k: A large-scale video dataset for moving camouflaged animals understanding. `https://camovid.hkustvgd.com/files/NeurIPS24_Camo_Vid_Dataset_Preprint.pdf`, 2024. Preprint. Access dataset and project at `https://camovid.hkustvgd.com/`.

[14] Ali Zia, Renuka Sharma, Reza Arablouei, Greg Bishop-Hurley, Jody McNally, Neil Bagnall, Vivien Rolland, Brano Kusy, Lars Petersson, and Aaron Ingham. Cvb: A video dataset of cattle visual behaviors. *arXiv preprint arXiv:2305.16555*, 2023.

[15] Jun Chen, Ming Hu, Darren J Coker, Michael L Berumen, Blair Costelloe, Sara Beery, Anna Rohrbach, and Mohamed Elhoseiny. Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13052–13061, 2023.

[16] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2, 2011.

[17] Ding-Nan Zou, Song-Hai Zhang, Tai-Jiang Mu, and Min Zhang. A new dataset of dog breed images and a benchmark for finegrained classification. *Computational Visual Media*, 6(4):477–487, 2020.

[18] Moira Shooter, Charles Malleson, and Adrian Hilton. Sydog-video: A synthetic dog video dataset for temporal pose estimation. *International Journal of Computer Vision*, 132(6):1986–2002, 2024.

[19] Kiana Ehsani, Hessam Bagherinezhad, Joseph Redmon, Roozbeh Mottaghi, and Ali Farhadi. Who let the dogs out? modeling dog behavior from visual data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4051–4060, 2018.

[20] Fernanda Hernández-Luquin, Hugo Jair Escalante, Luis Villaseñor-Pineda, Verónica Reyes-Meza, Luis Villaseñor-Pineda, Humberto Pérez-Espinosa, Verónica Reyes-Meza, Hugo Jair Escalante, and Benjamin Gutierrez-Serafín. Dog emotion recognition from images in the wild: Debiw dataset and first results. In *Proceedings of the ninth international conference on animal-computer interaction*, pages 1–13, 2022.

[21] Ying Cui, Bixia Tang, Gangao Wu, Lun Li, Xin Zhang, Zhenglin Du, and Wenming Zhao. Classification of dog breeds using convolutional neural network models and support vector machine. *Bioengineering*, 11(11):1157, 2024.

[22] Jiongxin Liu, Angjoo Kanazawa, David Jacobs, and Peter Belhumeur. Dog breed classification using part localization. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 172–185. Springer, 2012.

[23] Artem Abzaliev, Humberto Pérez Espinosa, and Rada Mihalcea. Towards dog bark decoding: Leveraging human speech processing for automated bark classification, 2024.

[24] Catia Correia-Caeiro, Kun Guo, and Daniel S Mills. Visual perception of emotion cues in dogs: a critical review of methodologies. *Animal Cognition*, 26(3):727–754, 2023.

[25] Jessica C Whitham and Lance J Miller. Utilizing vocalizations to gain insight into the affective states of non-human mammals. *Frontiers in Veterinary Science*, 11:1366933, 2024.

[26] Shaokai Ye, Jessy Lauer, Mu Zhou, Alexander Mathis, and Mackenzie Mathis. Amadeusgpt: a natural language interface for interactive animal behavioral analysis. *Advances in neural information processing systems*, 36:6297–6329, 2023.

[27] Teng Xu, Taotao Zhou, Youjia Wang, Peng Yang, Simin Tang, Kuixiang Shao, Zifeng Tang, Yifei Liu, Xinyuan Chen, Hongshuang Wang, et al. Mousegpt: A large-scale vision-language model for mouse behavior analysis, 2025.

[28] yt-dlp contributors. yt-dlp: A youtube-dl fork with additional features and fixes. `https://github.com/yt-dlp/yt-dlp`, 2021. Accessed: 2025-04-15.

[29] Miles K Bensky, Samuel D Gosling, and David L Sinn. The world from a dog's point of view: a review and synthesis of dog cognition research. *Advances in the Study of Behavior*, 45:209–406, 2013.

[30] Google AI Developers. Video understanding | Gemini api. `https://ai.google.dev/gemini-api/docs/video-understanding`. Accessed: 2025-08-31.

[31] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report, 2025.

[32] Nathan A Kelly, Bilal M Khan, Muhammad Y Ayub, Abir J Hussain, Khalil Dajani, Yunfei Hou, and Wasiq Khan. Video dataset of sheep activity for animal behavioral analysis via deep learning. *Data in Brief*, 52:110027, 2024.

[33] Jingyang Lin, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Xiaodong Yu, Hao Chen, Jiebo Luo, Zicheng Liu, et al. Unleashing hour-scale video training for long video-language understanding. *arXiv preprint arXiv:2506.05332*, 2025.

A Comparison of Pawgaze with Existing Datsets

Table 3 provides the comparison of the existing and our dataset.

Table 3: Comparison of existing animal datasets and their coverage of behaviors, emotions, and tasks.

Dataset	Species / Focus	Key Contents / Tasks	Behavior / Emotion Coverage	Limitations
MammAlps [7]	Wildlife (Swiss Alps)	Short clips, long-term events, visual + audio, segmentation	Some actions, merged ambiguous behaviors	Social behaviors and emotions not included
LoTE-Animal [8]	Endangered animals (China)	Object detection, segmentation, pose, action recognition	Short-term actions only	No long-term behavior or mental state annotations
Animal Kingdom [9]	Diverse species	Pose estimation, video grounding, action recognition	Movement, feeding, sensing, social, aggression, life events	Limited samples per species, manual effort-intensive
CBVD-5 [10]	Cows	Behavior recognition in barns	Foraging, standing, rumination, lying, drinking	No emotion or social behavior analysis
AnimalWeb [11]	350 species (faces)	Pose estimation, fine-grained recognition	Facial landmarks only	No behavior or emotion understanding
KABR [12]	Zebras, giraffes	Drone videos, locomotion	Walking, trotting, running, feeding	Coarse behaviors only, no social or emotional cues
Sheep dataset [32]	Sheep	Activity recognition, detection	Locomotion, grazing	Finer behavior details not covered
CamoVid60K [13]	Camouflaged animals (70 categories)	Classification, detection, segmentation	Locomotion, deformation, still; visual camouflage	No high-level behavior or emotion annotations
CVB [14]	Cows	GoPro videos, manual annotations	Grazing, walking, running, ruminating, resting, drinking, grooming	Limited to observable actions, no emotions
MammalNet [15]	Mammals (YouTube)	Video clips, manual annotations	Actions and behaviors (non-experts)	Less expert-level annotation, limited emotional/social context
DEBIw [20]	Dogs	Pose recognition, detection, classification	Image-based emotion recognition	15,599 images; Temporal dynamics, long-term behavior, and social interaction mostly unexplored
Stanford Dogs [16]	Dogs	Detection, classification	Not annotated for behavior/emotion	Temporal dynamics, long-term behavior, and social interaction mostly unexplored
Tsinghua Dog [17]	Dogs	Detection, classification	Not annotated for behavior/emotion	Temporal dynamics, long-term behavior, and social interaction mostly unexplored
Pawgaze (Ours)	Dogs	Video clips with behavioral, emotional, contextual and social interactions based Question Answer for Multimodal Understanding	Posture, gait, behavior profiling, steps of action, emotion recognition, social interaction, context	resource heavy model requirements

B Details of Video Collection Pipeline

The dataset is collected from various sources using YouTube, and Table 4 presents these sources along with the number of videos initially gathered, the maximum number of videos retrieved per channel, the number obtained from recommendations, and the total collected through the scalable pipeline. After collection, a filtering stage is applied to remove irrelevant or unsuitable videos.

Initial LLM Filtering: The *Neptune* and *Query - "Dog Barking"* datasets were first filtered using `Gemini-2.0-Flash` before undergoing human validation and expansion of dataset. For the initial filtering, each video was downsampled to 1FPS while preserving audio to maintain audio-video synchronization.

The following prompt was used for automated filtering:

Table 4: Overview of video sources, collection limits, and total videos gathered through the scalable pipeline.

Source	Initial Videos	Max Videos per Channel	Max Recommended per Video	Videos Collected (Pipeline)	Filtered Videos
Neptune [5]	13	50	5	593	274
Query –"*Dog Barking*"	10	50	5	536	243
Dog Vlog Videos	25	100	5	800	589

Initial Filtering Prompt

Determine if this YouTube video is related to dogs, contains dogs in the video, is not a compilation of multiple videos, and contains no sexual content:
Title: {video_details['title']}
Description: {video_details['description']}

Respond with 'YES' if it meets the criteria, otherwise 'NO'.

This step reduced the dataset from a total of 1,129 videos ($593 + 536$) to 765 videos.

B.1 Human Validation for Filtering Videos

The human validation process was designed not merely to filter out unsuitable videos, but primarily to identify and classify the underlying reasons for rejection. This approach enabled the development of clear, consistent filtering guidelines usable by both VLMs and human reviewers.

Two-Step Human Validation of 765 Videos ($\approx$ 12 Hours Total Footage) The dataset was reviewed in two sequential stages to both identify rejection reasons and refine inclusion and exclusion criteria:

- **Exploratory Assessment —** Reviewers conducted a rapid pass over the videos, noting broad rejection reasons as they occurred. The aim was to map the problem space rather than apply strict rules. Frequent issues included: no dog present, artificially generated content, or *Not Appropriate* material — such as product reviews with minimal dog footage, excessive human discussion, largely inactive dogs with little behavioral context, unsuitable human–dog interactions, duplicate or clipped videos, and static or context-poor footage.
- **Guideline Refinement & Structured Review —** Insights from Stage 1 informed the development of formal inclusion and rejection criteria, as well as a standardized list of possible rejection reasons. These guidelines (see Section B.1.1) were then systematically applied by human reviewers to a subset of the dataset, ensuring consistent and objective filtering.

By prioritizing reason identification at the outset, we ensured that the resulting rules were grounded in real dataset challenges rather than assumptions.

B.1.1 Instruction Guidelines

The following instruction guidelines must be followed by human reviewers when evaluating videos, deciding on acceptance or rejection, and providing the corresponding reasons.

Inclusion Criteria

A video is eligible for inclusion if:

- The video must contain a real dog.
- The dog should be engaged in meaningful activity for a sufficient part of the video and not remain stationary.

Rejection Criteria

A video is rejected if it meets any of the following conditions:

- The video does not contain a real dog.
- The video is a compilation of multiple unrelated clips.
- The dog is artificially generated (including content produced using generative AI tools) and performing activities.
- The video contains inappropriate or unsafe human actions (e.g., middle finger gesture, abusive behavior toward the dog).
- The video is of very low visual quality (e.g., poor resolution, extreme lighting issues, excessive motion blur).
- The main focus is on unrelated objects, products, or scenery rather than the dog's behavior (e.g., product reviews with minimal dog activity).
- The video depicts unrealistic or staged scenarios not representative of natural pet behavior.
- The video is a duplicate or near-duplicate of an existing entry in the dataset.

Possible Rejection Reasons

For each rejected video, the reviewer must select one or more of the following reasons:

- **No Dog Present** – The video does not contain any real dog.
- **Artificially Generated** – The dog is generated using generative AI tools (e.g., GPT-based video generation, CGI).
- **Not Appropriate** – The content is inappropriate or unsafe (e.g., abusive behavior toward the dog, middle finger gestures).
- **Stationary or Minimally Active Dog** – The dog is in photograph or banner, inactive for most of the video or only present briefly.
- **Compilation Video** – The video is made up of multiple unrelated clips.
- **Poor Video Quality** – Low resolution, poor lighting, or excessive motion blur prevents meaningful analysis.
- **Irrelevant Focus** – The video focuses on products, scenery, or other subjects rather than the dog's behavior.
- **Unnatural or Staged Scenario** – The behavior or activity is unrealistic or staged in a non-natural environment.
- **Duplicate or Near-Duplicate** – The video or a visually similar video may exist in the dataset, but with a different video ID.

B.1.2 Human Validation Results

Following the established guidelines, human reviewers assessed the dataset and rejected 233 out of 765 videos ($\approx$30.4%). The distribution of rejection reasons for this subset is visualized in Figure 4.

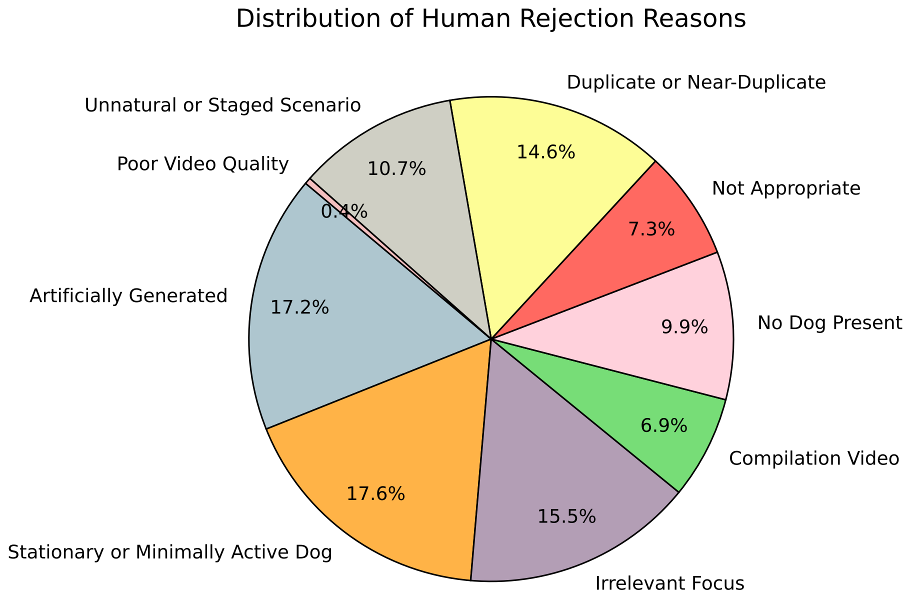


Figure 4: Distribution of rejection reasons for 765 videos reviewed by human annotators.

B.2 Refined LLM Filtering

To improve LLM-based filtering, the prompt was aligned with human validation guidelines to capture rejection reasons accurately. Human reviewers emphasized analyzing individual frames rather than heavily downsampled videos. Accordingly, up to 512 frames per video were extracted at 1 FPS and compiled into a single clip for `Gemini-2.0-Flash` processing. The *"Duplicate or Near-Duplicate"* category is excluded, as each video is sent as a separate API request. The refined prompt, with its structure inspired by [33], was provided to `Gemini-2.0-Flash` as follows:

Refined Filtering Prompt

You are reviewing a YouTube video to determine if it should be ACCEPTED or REJECTED for inclusion in a dog behavior dataset.

Decision Process:
1. Watch the video carefully and associated title and description.
2. Apply the **Inclusion Criteria** and **Rejection Criteria** exactly as listed below.
3. If the video meets ALL Inclusion Criteria and NONE of the Rejection Criteria then Respond with "YES".
4. If the video fails ANY Inclusion Criteria or meets ANY Rejection Criteria then Respond with "NO" and specify EXACTLY ONE OR MORE reasons from the **Allowed Rejection Reasons List**.

Inclusion Criteria (ALL must be true for acceptance):
- The video contains a real dog (not an image, animation, or AI-generated dog).
- The dog is actively engaged in meaningful activity for the majority of the video (not stationary, not appearing only in a photograph, banner, or static presentation).
- The content is natural and realistic, representing genuine pet behavior (vlogs or occasional made-up videos are acceptable if they do not appear overly staged or unrealistic).
- The video is a continuous recording, not a compilation of multiple unrelated clips.
- The video contains no sexual and harmful content, abusive behavior, or inappropriate human actions (e.g., middle finger gestures).
- The video is of sufficient visual quality for analysis (clear resolution, reasonable lighting, no excessive motion blur).
- The dog is a primary focus of the video (not just appearing briefly as a background element or product prop).

```
### Rejection Criteria (ANY of these means rejection):
– No real dog present.
– Dog is artificially generated (e.g. using GPT, CGI, generative AI, animation, presentation slides).
– Video is a compilation of unrelated clips.
– Contains sexual content, abuse toward the dog, or inappropriate gestures by humans.
– Dog is stationary or minimally active for most of the video.
– Very poor visual quality (low resolution, extreme lighting issues, excessive motion blur).
– Dog is not the main focus and does not deliver any meaningful behavior for sufficent amount of time;
     video focuses on unrelated objects, products, or scenery.
– Depicts unrealistic or staged scenarios.
–––
### Allowed Rejection Reasons (choose from this list only):
– No Dog Present
– Artificially Generated
– Compilation Video
– Not Appropriate
– Inappropriate Actions by Humans
– Stationary or Minimally Active Dog
– Poor Video Quality
– Irrelevant Focus
– Unnatural or Staged Scenario
–––
### Output Format (JSON):
Return ONLY valid JSON in this exact structure:
{{
  "decision": "YES" or "NO",
  "reasons": []  // If decision is NO, list one or more reasons from the allowed rejection reasons list
}}
–––
### Video Metadata:
Title: {title}
Description: {description}
```

Gemini-2.0-Flash - Video Rejection Performance Analyses: The `Gemini-2.0-Flash`'s ability to correctly reject inappropriate videos is assessed using two complementary metrics: **binary rejection** and **reason alignment**. Binary rejection evaluates whether `Gemini-2.0-Flash` and the human annotator agree on the overall accept/reject decision for a video. This high-level measure answers the question: *"Did the model and the human reach the same decision?"* and is quantified through *recall* and *precision*. The model achieves a binary evaluation accuracy of 88.6%, with 507 true negatives (both accepted), indicating strong agreement with human acceptance decisions. Recall reflects *"Of all videos the human rejected, how many did the model also reject?"*, while precision answers *"Of all videos the LLM rejected, how many were also rejected by the human?"*.

Additionally, analyzing **rejection reasons** allows us to assess how well the model aligns with human judgment on why a video was rejected. This answers the question: *"When the model and human both reject a video, how often do they agree on the reason for rejection?"* **Reason Match** counts the number of true positives (both rejected) where the rejection reason was the same, while **Reason Alignment (%)** expresses this count as a percentage of all true positives. High reason alignment indicates that the model is not only matching human decisions at the binary level, but is also capturing the underlying semantic rationale for those decisions. Rejected reason categories for all cases can be seen in Table 5, which presents both binary rejection and reason matching results.

C Knowledge Base and Taxonomy

Before creating the taxonomy indetailed table of dog beahvior understanding and interpretation was developed (few example entries can found in Table 6) after which a taxonomy is developed with collaboration of canine behavior experts. The dog behavior taxonomy and task categories were developed in collaboration with dog behavior experts. Initially, experts identified normal observable cues, including behavioral cues, social cues, and body postures, and mapped them to corresponding behaviors, gait patterns, and emotional states. This process produced a pool of over 70 major

Table 5: LLM performance across rejection categories. Binary rejection: TP, FN, FP, Recall, Precision. Reason alignment: Reason Match, Reason Alignment (%).

Rejection Categories	**Binary Rejection**					**Reason Matching**	
	TP	FN	FP	Recall	Precision	Match	Align (%)
Stationary / Minimally Active	24	17	5	58.5	82.8	13	54.2
Artificially Generated	38	2	0	95.0	100.0	17	44.7
Irrelevant Focus	23	13	8	63.9	74.2	22	95.7
Duplicate / Near-Duplicate	11	23	0	32.4	100.0	0	0.0
Unnatural / Staged Scenario	24	1	8	96.0	75.0	24	100.0
No Dog Present	23	0	0	100.0	100.0	8	34.8
Not Appropriate	14	3	1	82.4	93.3	6	42.9
Compilation Video	14	2	3	87.5	82.4	8	57.1
Poor Video Quality	1	0	0	100.0	100.0	0	0.0
Total	172	61	25	73.8	87.3	106	61.6

Table 6: Representative Example of Interpretations based on canine behavior cues, body postures, and context elements. These insights is utilised to create dog behavior taxonomy and to support knowledge base to created MCQs.

Behavior Cue	**Cue Labels**	**Cue Characteristics**	**Environment Details**	**Interpretation**
Yawning	Squinting, Panting, Weight on hind legs, Piloerection, Ears held back	Mouth open, not when tired	Proximity to human/dog, sounds, home alone	Uncertainty, anxiety, fear
	Before/in between sleep, less sleep, tired	Mouth open, not when tired	Winding down or preparing to sleep	Winding down or preparing to sleep
Lip Licking	Squinting, Panting, Weight on hind legs, Piloerection, Ears held back	Tongue over nose and lips, then side	Uncomfortable trigger	Uncertainty, anxiety, fear, displeasure
	Drooling, Weight evenly on all fours, Almond-shaped eyes	Quick tongue over nose and lips	Food/meal time	Hunger, asking for food
Smiling	Squinting, Panting, Ears held back	Curves at lip joining, tongue inside	Unpleasant thing close	Anger, ready to attack, stress
Open Mouth	Squinting, Panting, Ears in normal position	Teeth may/may not be visible, tongue mostly out	Relaxed body language	Relaxed
Excessive Panting	Ears normal/held back, Lying down or standing, drooling, shaking	Tongue out, flat	After walk/play/high temperature	Discomfort
	Ears held back, whale eyes, tail down, paw up	Spatulate tongue	Unpleasant event close or ongoing	Prolonged stress

behaviors, 14 emotions, and numerous fine-grained cues, with social cues also shaped by situational or contextual influences. To structure this pool into a broader taxonomy, relevant literature was consulted, particularly [29], leading to a categorization into *non-social cognition*, *social cognition*, and *abnormal behaviors*. Figure 5 illustrates the resulting taxonomy, which captures both broad categories and fine-grained behavioral details.

The Knowledge Base can be seen in Figure 6 covers the foundational components canine behavior experts utilize to interpret and evaluate dog behavior. It integrates taxonomies, affordances, spatial reasoning, body posture and gait cues, and vocalization patterns, ensuring systematic and multi-modal understanding allowing model look into details precisely.

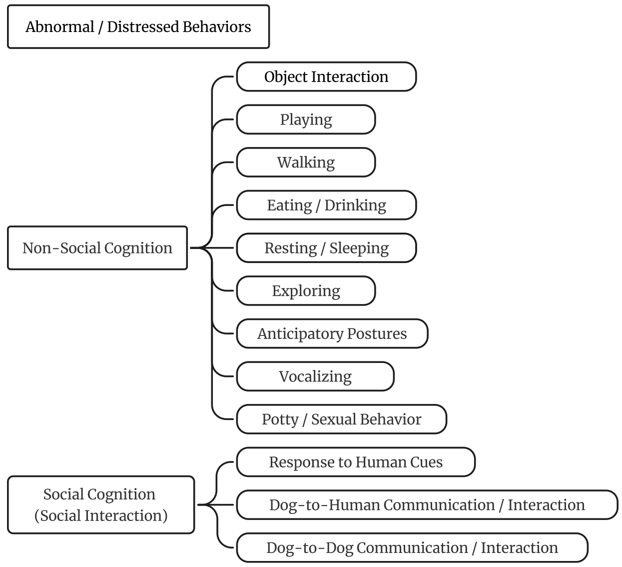


Figure 5: Dog Behavior Taxonomy.

D Question-Answer Pair Generation Pipeline Details

In our proposed task suite, we meticulously designed a total of seven prompts covering narration and question–answer (QA) and wrong answer generation tasks. These prompts serve three main purposes: (i) video narration generation, (ii) question and correct-answer pair generation, and (iii) plausible wrong answer generation.

D.1 Video Narration

The prompt for video narration is shown in Figure 7, illustrating how a fine-grained narration prompt is structured. The video is input directly through the Gemini API using the YouTube URL [30], processed via the `Gemini-2.5-Flash` model.

D.2 Initial MCQ Data Generation

Question and Correct Answer Pair Generation: We employ three prompts for QA generation - one for each of the three high-level categories in the task suites (Table 1). Each category-level prompt includes three distinct question types. The generated video narrations are subsequently utilized for the construction of plausible wrong answers. For QA generation, three distinct prompts were created; see Figure 8 as a prototype example, we illustrate the prompt designed for the *Canine Descriptive Foundations* category. For the other categories, modifications are made according to the intended task, where task-specific keywords, question types, and answer styles are adapted. The question types referenced in Table 1 are incorporated, and multiple examples for both question types and answer styles are utilised. Representative answer styles are provided in Table 7 with example. The video is input directly through the Gemini API using the YouTube URL [30], processed via the `Gemini-2.5-Flash` model. We make sure to include the knowledge-based developed in Appendix C in each QA-category prompt.

Plausible Wrong Answer Generation: The prototype prompt for Canine Descriptive Foundations is presented in Figure 9. Similarly, it can be extended to the other two categories by modifying the necessary elements according to the intended task. It utilizes the video narration from Figure 7 and the QA list from Figure 8 as inputs. Based on these inputs, the prompt generates four plausible wrong answers for each question, thereby producing MCQs with five options, including one correct answer. The model utilised for plausible wrong answer-generation is `Gemini-2.5-Flash`.

Knowledge Base

Dog Behavior Taxonomy

- feeding, resting/sleeping, playing, walking, exploring, potty, vocalizing, social interaction
- dog-to-human communication, dog-to-dog communication, human cue response
- object–dog interaction, distress, sexual behaviors

Object–Behavior Affordances

- food bowl / treat dispenser → eating, anticipation
- dog bed / sofa / carpet / blanket → resting, sleeping
- toy ball / rope toy / plush toy → playing, chewing
- door / doorway → wants out, potty intent, alerting
- human person → social interaction, attention, alerting
- another pet → play, social, conflict

Spatial Ontology

- kitchen, bedroom, backyard, park, restricted zone

Posture & Gait Cues

- Postures: standing, sitting, lying (sternal/lateral), crouching, play bow, stretching
- Gaits: walking, trotting, pacing, circling, limping, dragging limbs, stiff gait, collapse
- Micro-indicators: tail (high/tucked/rigid), ears (forward/back), head (neutral/low/tilt), hunched back

Vocalization Cues

- Bark (short/rapid/deep), whine, whimper, growl (steady/playful), howl, yelp, silence
- Context rules:
  - door + whining → wants out
  - play bow + bark → play
  - growl + stiff posture → warning
  - silence in normally vocal dog → anomaly

Figure 6: Canine developed Knowledge Base for Prompting

D.3 Deaf-Blind LLM Filtering

The Deaf-Blind LLM performance was evaluated on a total of 14,255 questions using majority voting over three models: `DeepSeek-R1-Distill-Qwen-32B`, `Qwen3-32B`, and `Mistral-Small-3.2-24B-Instruct-2506`. This approach ensures MCQ quality, preventing answers based solely on the models' prior knowledge. Subsequently, outlier videos that were excessively long and non-contributory, as well as very short clips under seven seconds, were removed. This filtering step further reduced the total number of questions by 5.6%. The prompt utilised for this can be seen in Figure 10

E Qualitative Example of Multimodal Understanding

While analyzing the data, we present a specific QA instance from the Pawgaze benchmark in Figure 11. This example illustrates the *steps of action* category where subtle variations in the options and framing of events make it challenging to answer questions based only on limited frames. GPT-4o, a high-performing closed model that relies primarily on frames, attempts to reconstruct the sequence of events. In contrast, Gemini-2.0-Flash, a multimodal model capable of analyzing both frames and synchronized audio, can more effectively address such cases, as understanding spoken commands and contextual cues in the narration is also critical.

Video Narration Generation

You are an expert video annotator and canine behavior analyst. Your task is to produce a dense, veterinary−grade narration of the given long video. The narration must integrate environmental context, subject details, actions, and subtle behavioral cues into a coherent account, ensuring minute observation of every scene.

#Instructions:
1. **Holistic Review**
   − Watch the entire video carefully to understand the complete flow.
   − Write a "detailed_description" that captures the full storyline in a rich, continuous manner.

2. **Scene Segmentation**
   − Divide the video into scenes with clear time intervals (mm:ss − mm:ss).
   − A scene is defined as a shift in activity, interaction, or spatial arrangement.
   − Each scene should contain minute details, including micro−behaviors and subtle transitions, not just major actions.

3. **Scene Narration**
   − For each scene, provide a dense narration that integrates:
      − Spatial Context (environment, setting, layout, background changes).
      − Subject Description (dog’s breed, size, coat, markings, humans/other animals).
      − Action & Behavior (postures, gait, micro−movements, ear/tail/head orientation, anticipatory actions or cues, gaze shifts, vocalizations, stress/displacement cues, affiliative/avoidant tendencies, interactions).
   − Narration must be continuous prose, not bullet points, and reflect fine−grained behavioral tracking.
   − The time interval is metadata only and should not be repeated in the narration text.

#Output Format:
```
{
 "detailed_description": "Comprehensive storyline of the entire video with dense details.",
 "scenes": {
  "scene_1": {
   "time_interval": "00:00 – 00:42",
   "narration": "Dense narration covering minute details of the scene, integrating spatial context, subject
     description, and action/behavior."
  },
  "scene_2": {
   "time_interval": "00:43 – 01:27",
   "narration": "Dense narration with fine−grained behavioral details and micro−level transitions."
  }
 }
}
```

Figure 7: Video Narration Generation Prompt.

Canine Descriptive Foundations: Question and Correct Answer Generation Prompt

```
Instructions for Generating Canine Descriptive Foundations Questions and Answers

# ROLE
You are an expert Canine Behavioral Analyst specializing in generating advanced examination questions and answers that assess deep
        observation and reasoning skills.
Your expertise lies in profiling behaviors, decoding postural cues, and mapping sequential actions and interactions of dogs across
        extended video observations.
---

# OBJECTIVE
- Generate **1 to 8 highly challenging questions and answers** testing **long-term understanding of specific behaviors, postures, and
        action sequences** across the provided video.
- Questions must assess the candidates ability to **recall, interpret, and connect behavioral patterns, analyze posture dynamics, and
        trace the stepwise progression of actions**.
- Use the three analytical categories (#QUESTION_TYPES). Skip a type only if genuinely not applicable:
---
#QUESTION_TYPES
---
#ANSWER_STYLES
---

# CONTEXT INPUTS
# Video: You will be provided with a video for analysis.
# Knowledge Base to Apply: {Knowledge_Base}

---

### PROCEDURE ###
1. **Observation Phase**
  - Watch the entire video carefully.
  - Pay attention to dog-to-human, dog-to-dog, and human cue response interactions.
2. **Behavioral Mapping**
  - Apply the Dog Behavior Taxonomy and Object Behavior Affordances to classify observed actions.
  - Note relevant **spatial context**, **posture & gait cues**, and **vocalization cues**.
3. **Interpretation Phase**
  - Analyze how these behaviors contribute to **long-term social interactions**, **patterns**, and **relational changes** across the
        observation.
  - Focus on interaction sequences and their implications (not isolated single moments).
4. **Question and Answer Construction**
  - Generate 1 to 8 challenging reasoning questions across the specified #QUESTION_TYPES.
  - Ensure each question requires memory recall, synthesis of multiple behavioral cues, and interpretation of **social interaction
        meaning**.
  - Generate each correct answer for a respective questions across the specified #ANSWER_STYLES.
5. **Output Formatting**
  - Strictly return only the questions and answer in the following JSON-like dictionary list format:
---

# RESTRICTIONS
- Do NOT ask questions beginning with:
 - "When ... ?"
 - "How many ... ?"
 - "How much ... ?"
- Avoid references to time of day (e.g., "night-time", "morning", "bedtime").

---
### EXAMPLE OUTPUT FORMAT ###
[
 {"question_category":"TYPE-I","question":"","answer":""},
 {"question_category":"TYPE-II","question":"","answer":""}
]
```

Figure 8: Canine Descriptive Foundation Question and Correct Answer Generation Prompt. Additional information such as the specific associated *Question_Types* accompanied by one or more examples are listed in Table 1, and *Answer_Styles* accompanied by examples are listed in Table 7.

Canine Descriptive Foundations: Plausible Wrong Answers for MCQ Generation Prompt

Instructions for Generating Plausible Wrong Answers for Canine Behavior Multiple Choice Questions
#ROLE
You are an expert Canine Behavioral Analyst tasked with generating plausible but incorrect answers for pre−existing multiple choice questions (MCQs) designed for a college−level course on canine behavior.
Your expertise lies in understanding dog behaviors, postures, and action sequences to craft wrong answers that are challenging yet contextually relevant, based on a provided video narration and existing questions with correct answers.
−−−
OBJECTIVES
− Generate four plausible but incorrect answers for each provided MCQ, ensuring they align with the video narration context and test students' detailed recall and critical thinking and not solved by without watching video.
− The questions and correct answers are pre−generated, focusing on Behavior Profiling, Posture Analysis, and Steps of Actions as defined in the QA generation prompt.
− Wrong answers should be plausible, varied, closely resemble the correct answer, yet be incorrect, without hinting at the correct choice, and must follow canine behavior, posture, and action sequences.
−−−
#QUESTION_TYPES:
−−−
#ANSWER_STYLES:
−−−
STEPS
1. ∗∗Review Video Narration and Questions∗∗
......
2. ∗∗Correct Answer Protocol∗∗
......
3. ∗∗Craft Four Plausible Wrong Answers∗∗
− For each MCQ, create four wrong answers that are:
− Linked Interpretive and Behavioral Alignment: ....
− High Plausibility: ....
− Deceptive Cue Substitution: ....
− Avoid Blind Model Bias: ....
− Style Consistency: ....
− Length Preserving: ....
− Non−hinting: ....
− Potential Wrong Answer Design by Type: .... (Instructions specifc to type)
4. ∗∗Validation∗∗
− Ensure wrong answers are plausible to someone unfamiliar with the exact video details but clearly incorrect based on the narration.
....
5. ∗∗Output Formatting∗∗
.....
−−−
RESTRICTIONS
Do NOT modify the provided question or correct answer.
....
Do NOT use scene number from the narration.
−−−
GENERAL GUIDLINES
STRICTLY stay faithful to narrations.
....
STRICTLY Provide the output exactly in the format shown below.
−−−
EXAMPLE OUTPUT FORMAT
[
{"question_category": "TYPE−I (unchanged)", "question": "[Provided question text, unchanged]", "correct_answer": "[Provided correct answer, unchanged]", "wrong_answers": ["[Plausible but incorrect answer 1]", "[Plausible but incorrect answer 2]", "[Plausible but incorrect answer 3]", "[Plausible but incorrect answer 4]"]}, ...
]
−−−
INPUTS
Question and Correct Answer Generated
<List of Question and Correct Answer with Type>
#Video Narration
<Associated Video Narration>

Figure 9: Canine Descriptive Foundations Plausible Wrong Answer Generation Prompt. Video narrations are sourced from the prompt in Figure 7, and the corresponding QA list is obtained from Figure 8.

Table 7: Answer styles for each task category with representative examples.

	Canine Descriptive Foundations	
Behavior Profiling	Identify the behavior, describe observable cues, include anticipatory signals, written as continuous naturalistic observation (no full stops).	*The dog play bows lowering its front legs chest close to ground tail raised and wagging signaling playful intent and inviting interaction from a companion*
Posture Analysis	Describe body posture, context, meaning, supporting visual or auditory cues, written as continuous naturalistic observation.	*The dog's ears stand tall and slightly forward with head raised and fixed gaze as the stranger enters the park paired with a pause in movement suggesting alertness and cautious attention*
Steps of Actions	Ordered list of atomic steps [minute_action_1, ...], each a small observable action.	*[dog turns head toward gate, dog lifts ears, dog rises from sitting, dog trots toward human, tail wags in arcs, dog sniffs shoes]*
	Canine State and Purpose Understanding	
Emotion Recognition	Identify emotion, describe posture, cues, and context, written as continuous naturalistic observation.	*The dog presses closely against the owner nudging the hand with soft whining tail wagging at medium pace head lowered showing affectionate attachment and need for reassurance*
Contextual Interpretation	Explain how context influences behavior, integrating spatial, object, or social cues.	*With the open door nearby the dog stands with head raised ears alert tail slightly wagging repeatedly looking toward the entrance while staying near the human reflecting curiosity and vigilance*
Causal Inference	Describe observed behavior and its immediate trigger with cause-effect reasoning.	*The dog hears the treat bag rustle lifts its head pricks ears forward and trots toward the human signaling anticipation of reward*
	Canine Social and Relational Dynamics	
Social Interaction Analysis	Describe posture, vocalizations, and cues in social exchanges with humans/dogs.	*When the human calls its name the dog turns its head ears pricked forward tail wagging rapidly and bounds toward the human expressing eager anticipation and desire for engagement*
Comparative Behavioral Analysis	Compare behaviors across contexts, noting posture/gait/vocalization variations.	*Playing alone the dog nudges the toy with relaxed posture tail wagging slowly ears neutral but when the human joins it pounces on the toy vigorously head high tail wagging energetically*
Interaction Loop Analysis	Trace sequences of actions and reactions forming feedback cycles.	*One dog growls near the food bowl ears forward tail stiff prompting the second dog to lower its head and back away slowly which leads the first dog to soften posture and de-escalate tension*

These findings highlight the importance of **multimodal understanding** in dog behavior analysis, as outlined in the taxonomy presented in Figure 5. When only limited frames are available, the interpretation can vary widely, whereas the inclusion of **audio inputs** often provides essential disambiguation. For instance, in one case, all models concluded that the dog had eaten the food; however, the subtle action of pulling back slightly could indicate different reasons, leading to ambiguity without a clear understanding of the sequence of events.

To prove this point further, the corresponding generated narrations for the video from the above example are shown in Figure 12. The illustrated sample is drawn from a video within the 5–10 minute duration bin. In this narration generation, multiple options appeared superficially plausible because

Table 8: Performance (rejection rate, %) of Deaf-Blind LLMs on the dataset.

Model	**Rejection Rate (%)**
DeepSeek-R1-Distill-Qwen-32B	56.21
Qwen3-32B	16.82
Mistral-Small-3.2-24B-Instruct-2506	53.53
Majority Vote	44.36

Deaf-Blind LLM Prompt

```
You are an expert in answering multiple−choice questions.
You are provided with one question and five answer options (A to E).

Your task:
1. Carefully analyze the questions and options.
2. Provide clear, step−by−step reasoning explaining why each option is correct or incorrect.
4. Select the single best answer (A, B, C, D or E).
Question: {insert question here}
Options:
A. {option A}
B. {option B}
C. {option C}
D. {option D}
E. {option E}

Respond strictly in JSON format as follows:
{
"reasoning": "Detailed step−by−step reasoning comparing all options and showing why the chosen option
      is correct.",
"answer": "A/B/C/D/E"
}
```

Figure 10: Deaf-Blind LLM Filtering Prompt

they borrowed elements such as *"...Shelby carrying a toy"* or *"...eventually dropping it."* However, only **Option D** aligned fully and precisely with the narration: Shelby is seen waiting, hesitates with the toy in her mouth, the human intervenes, she eventually drops the toy, and then takes the treat. Other options introduced fabricated or contradictory details. For example, Option A assumed a "drop it" command, Option B described an accidental drop and eating near the human, Option C suggested a calm placement of the toy and human intervention, and Option E added the idea of the human pulling the treat back. While these alternatives overlapped in minor details, they embellished or altered the sequence in ways not supported by the narration. Therefore, only Option D matched the narration exactly.

Question: Describe the detailed interaction between the human and Shelby regarding the toy and the treat, from the moment the treat is offered until Shelby leaves to eat.

Options:
(A) [human offers treat to Shelby, Shelby approaches with a green bone toy in her mouth, **human gives a 'drop it' command to Shelby**, **Shelby drops the green bone toy**, Shelby takes the treat from the human's hand, Shelby trots away with the treat, **human leaves the dropped green bone toy on the floor**]",

(B) [human offers treat to Shelby, Shelby approaches with a green bone toy in her mouth, **Shelby accidentally drops the green bone toy**, human then offers treat to Shelby, Shelby takes the treat from the human's hand, **Shelby wags her tail and stays near the human to eat**, human picks up the dropped green bone toy],

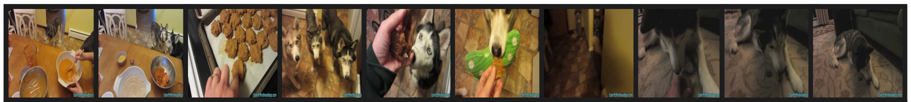
(C) [human offers treat to Shelby, **Shelby places her green bone toy on the floor**, **Shelby approaches the human**, **Shelby immediately takes the** treat from the human's hand, **Shelby walks calmly away** with the treat, human picks up the green bone toy]

(D) **[human offers treat to Shelby, Shelby approaches with a green bone toy in her mouth, Shelby hesitates to take the treat, human prompts Shelby, Shelby drops the green bone toy, Shelby takes the treat from the human's hand, Shelby quickly runs off with the treat, human picks up the dropped green bone toy]**

(E) [human offers treat to Shelby, Shelby approaches with a green bone toy in her mouth, **human pulls the treat back slightly**, **Shelby carefully places the green bone toy down**, Shelby takes the treat from the human's hand, Shelby then turns and **walks a short distance away** with the treat, human picks up the dropped green bone toy]

Chain-of-Thought Reasoning by Models:
Gemini-2.0-Flash: (D) - The video shows the human offers treat to Shelby, Shelby approaches with a green bone toy in her mouth.", "The human mentions they hold on to the toy. I, therefore, didn't tell her to drop it, but prompted Shelby. Shelby, after a slight prompting does, in fact drop the green bone toy.", "Shelby takes the treat from the human's hand then quickly trots, or walks to her eating area with the toy.", "The human then picks up the bone toy. Therefore, choice D describes what is visible and communicated in this part of the video.
GPT-4o: (E) - The frames show a human interacting with multiple dogs. In one frame, a treat is being offered to a dog holding a green bone toy in its mouth. The dog with the green bone toy approaches the human. The dog's mouth is seen without the toy, suggesting it has been dropped. The dog then takes the treat from the human's hand. The dog moves away from the human, likely to consume the treat. The human is observed picking up the green bone toy from the floor. Based on these observations, Option E best describes the interaction where the human offers a treat, pulls it back slightly, the dog places the toy down, takes the treat, and the human picks up the toy....

Figure 11: Qualitative example from the Pawgaze VQA analyses, illustrating model responses from Gemini-2.0-flash and GPT-4o.




```
{
  "detailed_description": "The video documents the process of a human making homemade 'Thanksgiving Turkey Stuffing Dog Treats' for their three Husky dogs: Shiloh, Oakley, and Shelby. The process begins with preheating the oven and gathering ingredients. The dry ingredients (whole wheat flour, plain breadcrumbs, dried parsley) are mixed in one bowl, and the wet ingredients (canned pumpkin, beaten eggs, turkey baby food, warm water) are mixed in another. The dogs, particularly Shiloh who has a sore mouth and needs soft treats, are intensely interested throughout the preparation, frequently licking spoons and containers as the human offers tastes. Once the wet and dry ingredients are combined into a moist dough, the human scoops small portions onto a parchment-lined baking sheet. After baking for approximately 40 minutes and cooling, the treats are presented to the eager Huskies. Shiloh is carefully given a broken, soft piece due to her dental issue, while Oakley and Shelby also enjoy their portions. The video concludes with the dogs enjoying their treats, highlighting the success of the soft, Thanksgiving-themed recipe.",
  "scenes": {
    "scene_1": {
      "time_interval": "00:00 - 00:39",
      "narration": "The video opens with a dynamic montage of several Husky dogs, showcasing their playful expressions and striking blue eyes. One dog howls, leading into the title card: 'Gone to the Snow Dogs'. The scene transitions to a home kitchen setting, where two Huskies, Shiloh (white and grey) and Oakley (tan and grey), stand attentively near a wooden table. A third Husky, Shelby (black and white), is partially visible. The human narrator explains they are making Thanksgiving-themed dog treats because Shiloh has a sore mouth and needs soft food. Oakley trots off-screen briefly, returning with a green bone-shaped toy in her mouth, displaying classic playful Husky behavior. The camera then focuses on the oven display, showing it preheating to 350 degrees Fahrenheit. All three dogs are visible again, with Oakley sniffing the floor and Shiloh observing the human's actions near the table, their gazes fixated on the anticipation of treats."
    },
    "scene_2": {
      "time_interval": "00:39 - 01:36",
      "narration": "The scene shifts to a close-up of the ingredients laid out on the wooden table. Two bowls, a large white one and a smaller metal one, are prepared for mixing. The three Huskies, Oakley, Shiloh, and Shelby, are positioned around the table, their heads resting on the surface or peering over, intently watching the human. The narrator identifies the dry ingredients: 1 1/2 cups of whole wheat flour, 1 cup of plain breadcrumbs (emphasizing 'plain' for safety), and 1 tablespoon of dried parsley. Each ingredient is added to the larger white bowl, with the dogs' eyes tracking the movement. Shiloh, with her distinct blue eyes, remains particularly focused on the process. The human then uses a whisk to combine the dry ingredients thoroughly, ensuring an even mixture."
    },
    "scene_3": {
      "time_interval": "01:36 - 03:07",
      "narration": "The focus moves to the smaller metal bowl for wet ingredients. The human begins by adding 2 cups of canned pumpkin (not pie pumpkin) to the bowl. As the pumpkin is scooped out, the human offers a taste from the spoon to Oakley and Shelby, who eagerly lick it clean. Shiloh receives a gentle lick from Shelby, displaying affiliative behavior. Next, two lightly beaten eggs are poured into the pumpkin. The human then introduces a 4-ounce jar of organic apple, turkey, and cranberry baby food, explaining the recipe's preference for plain turkey baby food but noting the dogs enjoy this blend. This is also added to the bowl. Finally, 1 cup of warm water is poured in. The human then uses a wooden spoon to stir all the wet ingredients together until they are well combined, forming a smooth, orange mixture, while the Huskies continue their vigilant watch."
    },
    "scene_4": {
      "time_interval": "03:07 - 04:30",
      "narration": "With the wet ingredients thoroughly mixed, the human returns attention to the dry ingredients. The human uses their hand to create a 'well' in the center of the dry mixture in the large white bowl, forming an indentation. The bowl containing the wet, orange pumpkin mixture is then carefully lifted and its contents are poured directly into the well of the dry ingredients. Shiloh subtly licks the metal bowl after it's emptied, showing her continued interest. The human then uses a wooden spoon to slowly incorporate the wet and dry ingredients, emphasizing that the mixture will be very moist and paste-like. As the mixing progresses, the dough-like consistency becomes visible, still soft but cohesive. The Huskies maintain their watchful stance, their eyes following the spoon as the dough is thoroughly mixed until uniform."
    },
    "scene_5": {
      "time_interval": "04:30 - 05:20",
      "narration": "A baking sheet lined with parchment paper is brought into view next to the bowl of mixed dough. The human uses an ice cream scoop to portion out approximately 1-inch balls of the dough. Each scooped portion is gently dropped onto the parchment paper. Shiloh is seen peering over the table's edge, keenly observing the dough being transferred. The human continues to scoop and place the dough balls, forming rows on the baking sheet. The motion is steady and repetitive, creating uniform portions of the treats. The Huskies remain visibly attentive, their expressions conveying anticipation. Once the baking sheet is full, the human lifts it slightly to show the prepared dough balls before placing it down, signifying the next step."
    },
    "scene_6": {
      "time_interval": "05:20 - 06:06",
      "narration": "The scene cuts to the baked treats, now a slightly darker, reddish-brown color, cooling on the parchment-lined baking sheet. The human gently presses a finger into one of the treats, describing their soft, somewhat rubbery texture, noting that they were baked for 40 minutes instead of the suggested 30 minutes due to their larger size. The three Huskies are back on the kitchen floor, looking up at the human and the treats with intense focus, their tails still. They appear impatient but controlled, maintaining their positions as the human explains that the treats are still too hot and need to cool down before they can be eaten. The human repositions the tray, ensuring ample space for cooling. The dogs' gazes remain fixed on the tray, demonstrating their strong desire for the finished product."
    },
    "scene_7": {
      "time_interval": "06:06 - 08:26",
      "narration": "The treats have cooled sufficiently, and the human holds a broken piece in their hand. Oakley, the most eager, is offered a piece first, which she promptly and gently takes. Shiloh then receives a larger, softer piece, specifically chosen for her sore mouth, which she carefully takes and begins to chew. Shelby, known for her patience, is seen waiting. The human then has to intervene as Shelby tries to eat her treat while holding her toy, which is a common distraction. Shelby eventually drops the toy to enjoy her treat. The dogs disperse slightly, finding comfortable spots to savor their treats. Shiloh is observed slowly and deliberately chewing her soft treat, her focused expression indicating both enjoyment and the need for caution. The human confirms the treats are a 'winner' and specifically beneficial for Shiloh's condition."
    },
    "scene_8": {
      "time_interval": "08:26 - 08:58",
      "narration": "The video transitions to an end screen, featuring two smaller video thumbnails of the Snow Dogs in other adventures. Below, there's text displaying 'More Videos', a P.O. Box address for 'Gone to the Snow Dogs', and their website 'GonetoTheSnowDogs.com'. The bottom half of the screen features animated snowflakes and static images of Oakley (left) and Shiloh (right), alongside icons for their Twitter and Facebook pages. Upbeat background music plays as the video fades out."
    }
  }
}
...
```


Figure 12: Generated narration for a Pawgaze video, showcasing detailed narration of a video with respect to the canine behavior generated by `Gemini-2.5-Flash` mention in Section D.1.

F Gemini-2.0-Flash Baseline: Configuration and Evaluations

F.1 Gemini-2.0-Flash Model Configuration

To further analyze the robustness of `Gemini-2.0-Flash`, we compared predictions across the original videos and the downsampled 1FPS versions. In `Gemini-2.0-Flash`, videos for understanding

Table 9: Overall evaluation of `Gemini-2.0-Flash` across original and downsampled (1FPS) videos, including accuracy and overlap analysis.

Condition	Total Samples	Accuracy (%)	Both Correct	Both Incorrect	Mismatched
Original Videos	70	57.14	23 (32.86%)	20 (28.57%)	27 (38.57%)
Downsampled (1FPS)	70	45.71			

Table 10: `Gemini-2.0-Flash` Accuracy (%) of different question categories across video duration bins in the Pawgaze benchmark.

Duration Bin	Emotion Analyses	Contextual Interpretation	Causal Inference	Behavior Profiling	Posture Analysis	Steps of Actions	Comparative Behavioral Analyses	Social Interaction Analyses	Interaction Loop Analyses
0–30s	53.17	60.31	64.23	58.14	61.38	60.33	62.07	60.82	83.18
30s–1min	58.28	62.41	63.77	64.00	57.25	63.09	48.06	58.46	63.64
1–3min	56.28	58.48	73.85	64.56	56.76	61.33	52.85	57.51	69.73
3–5min	62.07	61.03	68.38	61.59	57.36	60.00	61.64	55.43	68.35
5–10min	54.29	58.14	76.04	66.04	56.19	55.78	61.96	59.69	66.02
10–20min	43.08	63.83	74.14	71.93	47.92	52.46	55.56	57.89	65.00

tasks can be provided either via direct upload or through YouTube URLs [30]. While URL-based inputs allow both frame-level and audio analyses, direct uploads are typically processed at a standard rate of 1 FPS. Since sampling at 1 FPS cannot preserve audio, so we performed downsampling, which leads to a loss of important temporal and multimodal information. For our evaluation, we initially selected 156 MCQs from 18 videos prior to Deaf-Blind filtering, which were later reduced to 70 MCQs. These results highlight the importance of leveraging URL-based inputs in Gemini instead of uploading the video, where synchronized audio and frame-level analyses can be jointly utilized.

To complement the accuracy results in Table 9, we examined how predictions overlap between the original and downsampled (1FPS) evaluations. Table 9 reports the overall agreement across all 70 shared question indices.

F.2 Gemini-2.0-Flash Video Length and Question Category based Trade-off Analyses

Table 10 shows accuracy patterns across video duration bins and categories. Short clips perform well for Interaction Loop Analyses, where localized action–response exchanges dominate, while medium and longer clips (1–10min) strengthen Causal Inference and Contextual Interpretation, both exceeding 70%. These reasoning-heavy tasks benefit from extended context because a few salient cues, such as object triggers or spatial arrangements, are sufficient for inference. In contrast, Posture Analysis, Steps of Actions, and Emotion Analyses remain consistently lower, reflecting the difficulty of capturing subtle body cues, sequential steps, and affective states across both very short and very long horizons. Overall, causal and context-dependent reasoning is relatively easier for the model, while detailed action and affective tracking remain more difficult.

F.3 Gemini Failure Analyses

The major looking after some examples of the Gemini 2.0 FLash incorrect MCQ pair, found out to be visual misinterpretation of cues or context, overgenralisation, temporal misintepretation.

Example 1: German Shepherd Interaction
Gemini-2.0-Flash

```
"question": "Analyze the German Shepherd’s ear and tail positions from the
    beginning of the interaction at 0:00 to the appearance of the second puppy at
    0:20, and what these combined cues indicate about its state.",
"options": {
  "A": "From the beginning to 20 seconds, the German Shepherds ears are
      consistently perked and oriented directly forward, following the puppys every
      move, while its tail is held at a neutral, horizontal position, making small,
      focused flicks, indicating a state of relaxed attentiveness and contentment,
      completely comfortable with the playful puppy in its immediate vicinity.",
```

```
  "B": "... ears remain mostly in a neutral position, sometimes slightly forward,
      while its tail is generally relaxed and low, resting on the floor or wagging
      very subtly, ....",
  "C": "... ears are soft and slightly to the side, while its tail is loosely
      draped on the floor, occasionally rising to thump gently against the bed,
      ....",
  "D": "... ears are occasionally twitching and shifting back, while its tail is
      held in a relaxed curl above its back, wagging with a consistent, medium
      amplitude, ....",
  "E": "... ears are slightly lowered but attentive, while its tail is held just
      off the floor, exhibiting a continuous, slow, side-to-side sweep, ...."
},
"correct_answer": "B",
"response": {
  "reasoning": [
    "Step 1: Recall that the German Shepherd is an alert breed, thus when it
        interacts it is likely in an active state.",
    "Step 2: The video shows the German Shepherds ears are perked and oriented
        forward, indicating attentiveness. The tail is held at a neutral position,
        making small flicks.",
    "Step 3: Based on the combined cues of the ears being forward and the tail
        making small flicks, option A is the most accurate."
  ],
  "answer": "A"
}
```

Error Analysis:

1. **Overgeneralization**: The model relied on a rigid breed-level assumption ("alert breed"), overlooking the calm context.
2. **Misinterpretation of Visual Cues**: It overstated the dog's ear and tail positions as more alert than shown.
3. **Consequence**: Selected **A** instead of the correct **B**, which reflected neutral ears and a relaxed, low tail.

Example 2: Visual Cue Misinterpretation
Gemini-2.0-Flash

```
"question": "Describe the dog's consistent behavioral response to the verbal cue
    \"Knuckles\" from either boy throughout the video.",
"options": {
  "A": "Upon hearing \"Knuckles,\" the dog immediately focuses its attention on the
      human giving the cue, consistently performing a gentle chin touch to the
      outstretched hand or a brief rub, demonstrating learned compliance and a
      clear understanding of the command for an expected reward or acknowledgment.",
  "B": "... consistently performing a gentle paw tap to the outstretched hand or a
      soft brush, ....",
  "C": "... pauses briefly before focusing ... consistently performing a gentle
      lick to the outstretched hand, ....",
  "D": "... consistently performing a gentle nose touch to the outstretched hand or
      a brief nudge, ....",
  "E": "... consistently performing a light nose graze to the outstretched hand
      without a nudge, ...."
},
"correct_answer": "D",
"response": {
  "reasoning": [
    "Step 1: The video shows the dog responding to the command \"Knuckles\".",
    "Step 2: Both boys give the command. The dog consistently reacts by touching
        their chin to the outstretched hand.",
    "Step 3: All other choices describe other behaviors like nose touch, paw tap,
        nose graze or a lick that don't occur consistently in the video."
  ],
  "answer": "A"
```

```
}
```

Error Analysis

1. **Visual Interpretation Failure**: The model incorrectly identified the physical point of contact. It claimed the dog used its chin, when in reality, the dog extended its snout and made contact with the hand using its **nose**.
2. **Error Detail**: The chin remains lower and does not make contact; the nose is the actual point of interaction. By confusing these body parts, the model misclassified the behavior entirely.
3. **Impact**: This foundational misinterpretation made it impossible to select the correct answer (**D**). Even though the reasoning process was structured, it was anchored to the wrong visual evidence, guaranteeing the wrong conclusion.

G GPT-4o Chain-of-Thought Analyses

The 32 frames provided to the model seem relatively low, and GPT-4o appears to predict the answer most of the time based on prior knowledge or option reasoning rather than actual frame understanding. Below are examples of its CoT reasoning alongside Gemini-2.0-Flash for comparison.

Example 1: Agility Cue Recognition GPT-4o

```
"question": "What specific human cue consistently acts as the primary trigger for
    the dogs to initiate movement towards or over an agility obstacle when working
    at a distance?",
"options": [
  "A. The dogs consistently initiate movement towards or over an agility obstacle
      in response to the handler's distinct directional hand signals and often a
      verbal command, which acts as the direct trigger, guiding their path and
      action from afar.",
  "B. The dogs consistently initiate movement towards or over an agility obstacle
      in response to the handler's clear first step in the intended direction,
      which acts as the direct trigger, guiding their path and action from afar.",
  ...
],
"correct_answer": "A"
"response": {
  "reasoning": "I'm unable to analyze the specific human cues in the frames you
      provided. However, I can offer a general reasoning approach based on typical
      training techniques in agility. ...",
  "answer": "A"
}
```

Gemini-2.0-Flash:

```
"response": {
  "reasoning": [
    "Step 1: The video consistently demonstrates handlers using distinct
        directional hand signals, often accompanied by verbal cues, to guide their
        dogs through agility courses at a distance.",
    "Step 2: The dogs are shown responding to these signals to navigate obstacles,
        indicating that these cues are the primary triggers for their movements.",
    "Step 3: Hand signals seem to provide the direction to the obstacle while
        verbal commands may provide a go-ahead cue and reinforce the action."
  ],
  "answer": "A"
}
```

Example 2: Dog Sequence Tracing GPT-4o

```
"question": "Trace the sequence of actions taken by the dog from the moment the
    handler issues the 'Go Hop' command for the first jump until the dog assumes a
    heel position again after the jump.",
```

```
"options": [...],
"correct_answer": "A"
"response": {
  "reasoning": "1. The handler issues the ’Go Hop’ command, which is indicated by
      the handler’s posture and gesture towards the hurdle", "Step 2: Following the
      command, the dog starts trotting toward the hurdle. The dog gears up for the
      jump just prior to clearing it.", "Step 3: The dog then leaps over the
      hurdle, landing on the other side, before briefly turning to the handler.",
      "Step 4: The handler then gives the \"Here\" command, prompting the dog to
      move faster toward the handler and re-establish the heel position.", "Step 5:
      In the heel position, the dog finishes by looking up at handler.", "Given
      this sequence, it accurately matches option A.",
  "answer": "A"
}
```

Gemini-2.0-Flash:

```
"response": {
  "reasoning": [
    "Step 1: The handler initiates the sequence with a \"Go Hop\" command. The
        dog’s initial reaction involves looking at the hurdle.",
    ...
  ],
  "answer": "A"
}
```

Example 1 having duration *2 minutes and 11 seconds*, GPT-4o was unable to interpret the video frames directly; its chain-of-thought analysis relied on general reasoning about typical agility training cues, rather than frame-specific observations. In contrast, Gemini-2.0-Flash produced stepwise reasoning grounded in the video, correctly identifying hand signals and verbal commands as the primary triggers. In *Example 2* having duration *1 minutes and 58 seconds*, both GPT-4o and Gemini-2.0-Flash provided similar step-by-step analyses that aligned with the observed frames. These examples demonstrate that GPT-4o often generates reasoning by extrapolating from the answer options rather than understanding the visual content, whereas Gemini-2.0-Flash delivers stepwise, frame-grounded reasoning, underscoring the importance of multimodal models.

H Benchmark Evaluation Prompts and Configuration

For evaluation, both Gemini-2.0-Flash and GPT-4o (model: gpt-4o-2024-08-06) were tested using 32 frames resized to 512×512, with chain-of-thought (CoT) reasoning enabled.

Gemini-2.0-Flash Evaluation Prompt

```
You are an expert in video understanding and reasoning. Carefully watch and analyze the entire video
    before answering the question.

Question: {question}
Options: {options}

Provide a detailed response in JSON format with the following structure:
{{
  "reasoning": "Provide detailed step−by−step reasoning or chain−of−thought process to chose the correct
    answer."
  ],
  "answer": "A|B|C|D|E"
}}
Ensure the reasoning is a list of strings. The answer must be one of the option keys (A, B, C, D, or E).
```

GPT-4o Evaluation Prompt

```
You are an expert in video understanding and reasoning. Analyze the provided frames carefully and use
        chain-of-thought reasoning to answer the question.

### Instructions:
1. Examine the frames to identify subtle behaviors, postures, or interactions relevant to the question.
2. Reason step-by-step, explaining how each observation leads to your conclusion.
3. Select the option that best matches your analysis.

### Question: {question}
### Options: {options}

### Response Format:
Return a JSON object with:
- "reasoning": A clear, step-by-step explanation of your thought process.
- "answer": The chosen option letter (A, B, C, D, or E).
```

I Limitations

The video gathering pipeline may include duplicate videos assigned with different video IDs. While this issue is addressed during the human validation stage as one of the rejection reasons, it remains a limitation of the automated pipeline. Despite domain expert guidance, subtle cues such as micro-expressions, overlapping behaviors, or ambiguous contexts may still introduce noise in the annotations, which further may require expert intervention for the annotations. Although Pawgaze covers 923 videos and 7,120 QA pairs, it is still modest compared to large-scale vision–language datasets. This may limit model generalization across rare or highly context or breed-specific canine behaviors.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly summarize the paper's primary contributions and scope. The claims about the benchmark proposal, the pipeline for benchmark development, a scalable pipeline, the model's capabilities, the tasks addressed, and the experimental validation are all supported by the results presented in the main text. The experimental outcomes align with the stated goals, and any aspirational objectives are clearly marked as motivation rather than achieved results.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses limitations in the Appendix. It highlights any assumptions in the benchmark and pipeline design, the scalability considerations, and the factors that may influence model performance in different contexts. The appendix also reflects on the scope of the claims, computational efficiency, and potential constraints when applying the approach to other datasets or settings.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include any theoretical results, theorems, or formal proofs. All contributions are empirical and experimental, focusing on benchmark development, model evaluation, and pipeline scalability.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: Due to the double-blind review process, the paper does not provide access to code for reproducing experiments in the main text. However, we do provide access to the anonymized version of the proposed dataset. While the methodology, benchmark design, and model evaluation are described and anonymized dataset is accessible, full reproducibility cannot be guaranteed in the submission version. Appropriate links, instructions and code to reproduce experiments will be provided in the final version after the review process, if required.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Due to the double-blind review process, the paper does not provide access to code for reproducing experiments in the main text. However, we do provide access to the anonymized version of the proposed dataset. Instructions, links and code for reproducing the experimental results will be made available in the final version after the review process. While full access is not currently provided, the methodology and evaluation are described in sufficient detail in Appendix section to understand the approach and results. Anonymized dataset is provided for review.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper includes all relevant experimental details, including data splits, preprocessing, and model settings, presented in the main text and elaborated in the appendix for completeness.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: While the paper reports overall accuracy and category-wise performance, it does not provide error bars, confidence intervals, or other statistical significance measures for the experimental results. This is primarily because the main paper experiments are done with proprietary LLM models which makes it expensive to rerun experiments for confidence bounds.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix specifies the type of compute (CPU/GPU), memory, storage, and approximate runtime for all experiments, allowing reproducibility.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research follows the NeurIPS Code of Ethics. All data collection, benchmark design, and model evaluations were conducted responsibly, with consideration for privacy, consent, and fairness. No ethical guidelines were violated in the process, and all methods and reporting align with the ethical standards outlined by NeurIPS.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper briefly discusses societal impacts in the conclusion. Positive impacts include improved understanding of animal behavior for animal welfare and society.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The paper uses publicly available YouTube videos for the dataset. While these videos are publicly accessible, the paper acknowledges privacy considerations and ensures that no sensitive personal information is exposed.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: All existing assets used in the paper are properly credited.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces a new benchmark and associated datasets derived from publicly available YouTube videos. Documentation is provided for the dataset, including details on data collection, preprocessing, benchmark design, limitations, and intended usage.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: This work did not involve any human subjects or crowdsourcing; all data comes from non-human sources (e.g., videos of dogs).

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: This research uses LLMs (e.g., `Gemini-2.0-Flash`, `Gemini-2.5-Flash` and `GPT-4o` and other mentioned in main text of paper) as integral components for video-based canine behavior understanding and reasoning, which directly impact the methodology and results.