AI for Animal Pain Assessment: a New Challenge for Bioacoustics

George Martvel
University of Haifa
martvelge@gmail.com

Annika Bremhorst
University of Bern

Mayara Travalini de Lima
São Paulo State University

Stelio Pacca Loureiro Luna
São Paulo State University

Anna Zamansky
University of Haifa

Abstract

We introduce CoViNLE, a coarse-to-fine architecture for audio-visual question answering that refines global and local cues in videos into natural language descriptions. We evaluate its effectiveness in assessing acute pain in canines, which is challenging due to the subtle behavioural and bioacoustic signals animals display. CoViNLE is tested against veterinary expert scores using the Glasgow Composite Pain Measure Index. The results reveal significant limitations: vision-only models overlook important behaviours, while audio-based models, such as fine-tuned Whisper and NatureLM-audio, identify vocal pain indicators but often produce unstable results and hallucinations. This highlights the need for more robust audio models and more diverse datasets for training bioacoustic large language models.

1 Introduction

Many Multimodal Large Language Models (MLLMs) now accept video data; however, this modality is less developed compared to image data. In such models, the input is split into a visual stream (sampled frames) and an audio stream. Each stream is encoded with a modality-specific backbone to produce token sequences, which are then projected/fused (via concatenation or cross-attention) for joint reasoning [1–3]. Thus, the choice and quality of these encoders directly affect end-to-end performance. The visual capabilities of MLLMs have advanced rapidly: benchmarks like MMMU demonstrate progress on heterogeneous image types, and recent models report strong results across various vision tasks [4, 5]. By contrast, audio understanding lags behind: audio-language benchmarks highlight substantial headroom on non-speech audio and common failure modes (e.g., hallucination, temporal-order mistakes) [6–8]. Widely used environmental sound datasets remain relatively limited; for example, ESC-50 comprises 50 classes, and UrbanSound/UrbanSound8K covers only 10 classes. At the same time, even large-scale AudioSet provides weak clip-level labels (sound presence) across 632 classes, limiting fine-grained supervision [9–11].

Animals express their internal states through a wide range of vocalisations across species. Their structure encompasses intensity, pitch, spectral/timbre features, and temporal traits including duration and call rate [12–14]. Beyond graded variation, some species produce functionally referential call types (e.g., distinct alarm calls for different predators), evidencing semantic specificity [15, 16]; information may also be encoded across call sequences rather than single calls [17]. Capturing this acoustic variety is essential for studying animal behaviour, communication, and cognition [18]. In practice, large parts of wildlife-audio workflows still rely on manual inspection and labelling, and annotation cost and availability remain bottlenecks [18, 19]. When machines are integrated into the process, general-purpose audio models often perform well on broad categories but struggle

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

with fine-grained, context-dependent bioacoustic classifications and precise temporal localisation, a challenge aggravated by weak clip-level labels in datasets such as AudioSet [20]. Recent evaluations indicate that generic audio-language models struggle with species-level or non-speech bioacoustics without specialised training [21], highlighting a form of "deafness" beyond very general concepts.

In this study, we propose a CoViNLE algorithm (**Co**arse-to-fine **Vi**deo **N**atural **L**anguage **E**ncoding), an LLM-based architecture for audio-visual question answering (AVQA) that produces human-interpretable natural language descriptions and justifications. As a pilot study, we evaluated the algorithm using video recordings of dogs taken during clinical pain assessments at a veterinary clinic. We compared pain scores from an expert with those generated by the algorithm, based on a common clinical pain assessment scale in veterinary medicine (CMPS-SF [22]). We report preliminary results and an error analysis highlighting strengths and limitations of current MLLMs, including GPT models [23], NatureLM-audio [24], and Whisper [25, 26]. Our goal is to lay the foundation for an agentic, accurate AVQA system capable of interpreting animal visual and acoustic signals, thereby advancing our understanding of animal communication through AI.

2 Methods

2.1 Dataset

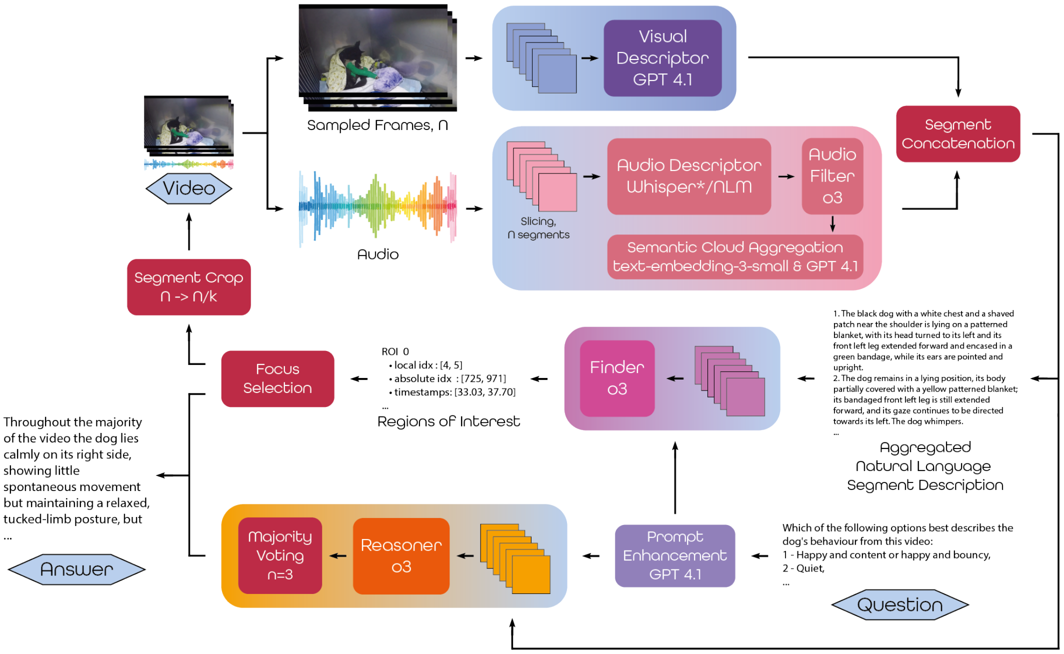
The dataset used in this study comprises 184 videos, featuring dogs before and after a surgical procedure during the pain assessment process. These videos are used by experts to score pain using the Glasgow Composite Pain Measure Index (CMPS-SF [22]). Each question in the pain scale questionnaire has categorical answer options, with the selected option counting as the question score. An experienced veterinarian who had completed a Residence and an MSc in Veterinary Anesthesiology scored the videos, resulting in five question-answer pairs and a cumulative pain score. If the total score exceeds 5, it indicates pain and suggests the need for analgesics. From the dataset, we selected a subset of 36 videos: 18 with the lowest pain scores (0 and 1) and 18 with the highest pain scores (9 to 12), to represent the most distinct examples of each category (see details in the Supplementary Materials). We also adapted the original questions from the questionnaire to improve their suitability as input for LLMs (provided in the Supplementary Materials). The resulting questions have the following themes: *Vocalisation* — how the dog vocalises; *Painful area* — how the dog interacts with the painful area; *Human interaction* — how the dog reacts to the veterinarian touching the painful area; *Behaviour* — general dog's behaviour; *Condition* — general dog's condition.

2.2 Pipeline

The CoViNLE pipeline's architecture is shown in Figure 1. The algorithm is coarse-to-fine, meaning that the answer-finding process is iterative, and on each step, a smaller video segment is analysed. In the current study, we chose the initial frame sampling and audio slicing parameter $N_0 = 20$ and limited the number of recursive steps to three due to resource constraints. Since most of the models in the pipeline are prompt-dependent, we provide corresponding prompts in the Supplementary Materials. All experiments were performed using Google Colab service and an A100 GPU (40 GB GPU RAM, 83.5 GB System RAM).

Visual Encoder. The Visual Encoder module includes the Visual Descriptor model (GPT 4.1), which generates natural-language descriptions of the N input frames. The model is instructed to focus solely on visual information and to avoid using subjective evaluative adjectives (e.g. sad, happy, gentle).

Audio Encoder. The Audio Encoder module comprises the Audio Descriptor model, the Audio Filter model, and a Semantic Cloud Aggregation submodule. The Audio Descriptor model (finetuned Whisper [26] (*small* and *large*) or NatureLM-audio [24]) takes N audio slices as input, and produces a one-sentence caption for each slice. Resulting captions often contain hallucinations and irrelevant information (e.g., misinterpretations of animal sounds or descriptions of background noise), so they are filtered with an Audio Filter model (o3), which "interprets" them as most probable based on context. However, those filtered descriptions are not stable, meaning that the Audio Filter model may produce different captions for the same input caption (e.g. input: *"A bird squawks."*, output: *"No information."*/*"A dog barks."*). To mitigate the instability, the results of several Audio Filter model runs are processed by the Semantic Cloud Aggregation submodule. In this submodule, $n_{SCA} = 10$ sampled captions are encoded by a Text Encoder model (OpenAI's text-embedding-3-small). The



ROI 0
- local idx : [4, 5]
- absolute idx : [725, 971]
- timestamps: [33.03, 37.70]

...

1. The black dog with a white chest and a shaved patch near the shoulder is lying on a patterned blanket, with its head turned to its left and its front left leg extended forward and encased in a green bandage, while its ears are pointed and upright.
2. The dog remains in a lying position, its body partially covered with a yellow patterned blanket; its bandaged front left leg is still extended forward, and its gaze continues to be directed towards its left. The dog whimpers.

...

Which of the following options best describes the dog's behaviour from this video:
1 - Happy and content or happy and bouncy,
2 - Quiet,
...

Throughout the majority of the video the dog lies calmly on its right side, showing little spontaneous movement but maintaining a relaxed, tucked-limb posture, but
...

Figure 1: The CoViNLE architecture processes input video by separating it into two streams — sampled frames and audio. Each stream is encoded, and the descriptions are merged segment-wise. The combined natural language description is sent to the Finder and Reasoner modules with a modified question prompt. If the models determine that the description contains a final answer, it is provided. If not, the video is cropped based on regions of interest identified by the Finder module, and the process iterates until an answer is produced or the maximum number of iterations is reached. The Whisper* model is a finetuned model published by Kadlčík et al. [26]

resulting embeddings are compared with cosine similarity across captions to form a similarity matrix and clustered by thresholding similarity at $\tau = 0.85$. If the largest cluster does not cover at least half of the available candidates, "No information" is assigned to that segment. Otherwise, sentences in the largest cluster are merged by an Audio Caption Aggregator model (GPT 4.1) into the final audio segment caption.

Finder. The Finder module consists of the same-name model (o3) and takes as input a merged audio-visual segment description from encoder modules with an enhanced prompt (question edited by a Prompt Enhancement GPT 4.1 model). It then identifies the regions of interest, which potentially may contain information relevant to the question. If regions of interest are close to each other (distance threshold is set as $\frac{N}{5}$), they are merged. The output of this module is a segment list, containing indexes and timestamps of segment start/end.

Reasoner. The Reasoner module consists of the aggregation of Reasoner models (o3) by the majority voting mechanism ($n = 3$) to increase the answer stability. It takes a merged audio-visual segment description from encoder modules with an enhanced prompt as input and produces an answer in natural language. As an option, it may also produce a numeric answer.

Focus Selection Function. This function determines if the output produced by the Reasoner module is the final answer to the input question. The decision is positive if the Finder's output is empty, or the length of one of the regions of interest is larger than half of the total number of segments (half of the video). If the decision is negative, the input video is cropped by the selected region's of interest timecodes, and the number N of sampled frames is reduced by $k = 1.5$ to avoid dense sampling.

Answer Selection. During the coarse-to-fine process, the algorithm may produce multiple regions of interest on each step, resulting in a "tree of answers" (one answer for each final region). The final answer is selected as the maximum numeric value among all the nodes. Such a choice is task-specific (the answer options in the CMPS-SF are ranked by severity) and may vary in other tasks.

2.3 Metrics

For each of the five CMPS-SF questions, we compared the predicted score to the veterinarian-placed score and reported two per-variable metrics: (i) *accuracy*, the proportion of exact matches; and (ii) *mean absolute error* (MAE), the average absolute difference between scores. For a total score, we added the five predicted values and the five ground-truth values, labelling a video as positive if its total score is ≥ 5 (negative otherwise), and then compared the resulting labels. From this comparison, we computed *binary accuracy* and the *F1-score*.

3 Preliminary Results

We present the classification results for each question, along with a binary classification result for the total pain score in Table 1. Our findings indicate that the visual-only pipeline was unable to comprehend animal behavioural patterns, highlighting the necessity of the audio modality. However, the pipelines utilising audio models still exhibited high error rates. Notably, the pipeline that employed the finetuned Whisper [26] model generally outperformed the pipeline that used the NatureLM-audio [24]. Additionally, our proposed pipeline struggled with evaluative tasks, which involve assessing and interpreting the dog's behaviour and condition, and performed better on factual questions.

4 Discussion

In this study, we introduce a novel Coarse-to-Fine Video Natural Language Encoding (CoViNLE) zero-shot architecture designed for the audio-visual question answering task. As a proof of concept, we demonstrate its performance using a dataset of recordings of dogs undergoing pain assessment procedures with the CMPS-SF pain scale. The primary advantage of our proposed architecture is the "transparency" of its encodings, which is particularly beneficial in the animal domain, where understanding the reasoning behind decisions is critical. In the task of behavioural pain assessment, natural language descriptions generated by large language models serve as interpretable representations of video data, providing valuable information to veterinarians. Furthermore, the coarse-to-fine approach of our architecture addresses the traditional challenges associated with frame sampling rates, allowing it to capture information that may occur "between the frames".

The results obtained highlight the current limitations of existing MLLMs in capturing animal-specific auditory information, which is essential for assessing complex internal states. While the finetuned

Table 1: CoViNLE performance summary with different Audio Descriptor backbones.

	WhisperSmall	WhisperLarge	NatureLM	None
Total Score Metrics				
Accuracy	0.67	**0.69**	0.53	0.50
F1	**0.73**	0.72	0.60	0.00
Accuracy per question				
Vocalisation	0.44	**0.53**	0.31	0.25
Painful Area	0.86	**0.92**	**0.92**	0.03
Human Interaction	**0.58**	0.53	0.56	0.00
Behaviour	0.39	**0.42**	0.28	0.31
Condition	0.25	**0.36**	0.17	0.19
Average	0.51	**0.55**	0.44	0.16
MAE per question				
Vocalisation	**0.75**	0.78	0.97	0.78
Painful Area	0.17	**0.11**	**0.11**	0.97
Human Interaction	**1.25**	1.64	1.69	1.97
Behaviour	1.33	**1.17**	1.56	1.42
Condition	1.50	**1.36**	1.69	1.47
Average	**1.00**	1.01	1.21	1.32

Whisper and NatureLM-audio models are capable of processing general animal audio, we have observed high rates of instability and hallucination in real-world clinic scenarios. In our experiments, we observed that the absence or misinterpretation of the signal from the Audio Encoder often drastically impacts performance overall, causing false negative predictions (not detecting important regions) from the Finder module. This highlights the necessity of the development of more robust and perceptive bioacoustic models. In the CoViNLE pipeline, we have implemented measures such as filtering and embedding aggregation to address these issues, but there is still significant potential for improvement.

Since this work is only a preliminary proof-of-concept study, we acknowledge that a more detailed investigation is needed to properly test the vision and audio modules of the proposed algorithm. In future work, we plan to conduct an ablation study focusing on prompts, model backbones, and hyperparameters, investigate and mitigate the major causes of errors, and apply our approach to other datasets to enhance its generalizability and robustness.

Acknowledgments and Disclosure of Funding

The authors have no funding, financial support, or acknowledgements to declare.

References

[1] Sun, G., Yu, W., Tang, C., Chen, X., Tan, T., Li, W., ..., and Zhang, C. (2024). video-salmonn: Speech-enhanced audio-visual large language models. arXiv preprint arXiv:2406.15704.

[2] Chowdhury, S., Nag, S., Dasgupta, S., Chen, J., Elhoseiny, M., Gao, R., and Manocha, D. (2024, September). Meerkat: Audio-visual large language model for grounding in space and time. In *European Conference on Computer Vision* (pp. 52-70). Cham: Springer Nature Switzerland.

[3] Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., ..., and Bing, L. (2024). Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476.

[4] Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., ..., and Chen, W. (2024). Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9556-9567).

[5] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., ..., and Lin, J. (2024). Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191.

[6] Yang, Q., Xu, J., Liu, W., Chu, Y., Jiang, Z., Zhou, X., ..., and Zhou, J. (2024). Air-bench: Benchmarking large audio-language models via generative comprehension. arXiv preprint arXiv:2402.07729.

[7] Wang, B., Zou, X., Lin, G., Sun, S., Liu, Z., Zhang, W., ..., and Chen, N. F. (2024). Audiobench: A universal benchmark for audio large language models. arXiv preprint arXiv:2406.16020.

[8] Kuan, C. Y., & Lee, H. Y. (2025, April). Can large audio-language models truly hear? tackling hallucinations with multi-task assessment and stepwise audio reasoning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

[9] Piczak, K. J. (2015, October). ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 1015-1018).

[10] Salamon, J., Jacoby, C., and Bello, J. P. (2014, November). A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 1041-1044).

[11] Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., ..., and Ritter, M. (2017, March). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 776-780). IEEE.

[12] Taylor, A. M. and Reby, D. (2010). The contribution of source–filter theory to mammal vocal communication research. *Journal of Zoology*, 280(3), 221-236.

[13] Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *The Journal of the Acoustical Society of America*, 102(2), 1213-1222.

[14] Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. **The American Naturalist**, 111(981), 855-869.

[15] Seyfarth, R. M., Cheney, D. L., and Marler, P. (1980). Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. *Science*, 210(4471), 801-803.

[16] Townsend, S. W. and Manser, M. B. (2013). Functionally referential communication in mammals: the past, present and the future. *Ethology*, 119(1), 1-11.

[17] Kershenbaum, A., Blumstein, D. T., Roch, M. A., Akçay, Ç., Backus, G., Bee, M. A., ..., and Zamora-Gutierrez, V. (2016). Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biological Reviews*, 91(1), 13-52.

[18] Stowell, D. (2022). Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10, e13152.

[19] Stowell, D., Wood, M., Stylianou, Y., and Glotin, H. (2016, September). Bird detection in audio: a survey and a challenge. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1-6). IEEE.

[20] Kong, Q., Yu, C., Xu, Y., Iqbal, T., Wang, W., and Plumbley, M. D. (2019). Weakly labelled audioset tagging with attention neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11), 1791-1802.

[21] Miao, Z., Elizalde, B., Deshmukh, S., Kitzes, J., Wang, H., Dodhia, R., and Ferres, J. L. (2025). Multi-modal Language models in bioacoustics with zero-shot transfer: a case study. Scientific Reports, 15(1), 7242.

[22] Reid, J., Nolan, A. M., Hughes, J. M. L., Lascelles, D., Pawson, P., and Scott, E. M. (2007). Development of the short-form Glasgow Composite Measure Pain Scale (CMPS-SF) and derivation of an analgesic intervention score. *Animal welfare*, 16(S1), 97-104.

[23] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ..., and McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

[24] Robinson, D., Miron, M., Hagiwara, M., Weck, B., Keen, S., Alizadeh, M., ... and Pietquin, O. (2024). Naturelm-audio: an audio-language foundation model for bioacoustics. arXiv preprint arXiv:2411.07186.

[25] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492-28518). PMLR.

[26] Kadlčík, M., Hájek, A., Kieslich, J., and Winiecki, R. (2023). A whisper transformer for audio captioning trained with synthetic captions and transfer learning. arXiv preprint arXiv:2305.09690.

A Technical Appendices and Supplementary Material

Dataset. The video dataset is available at `https://osf.io/re6pw/overview?view_only=3a047fd9032a4c0e967a769c70e8b122`. The study and data collection were approved by the Animal Use Ethics Committee (CEUA 0053/2021) of the Faculty of Veterinary Medicine and Animal Science (FMVZ) of the São Paulo State University (UNESP).

Questions from the CMPS-SF questionnaire (adapted for prompt usage).

1. Which of the following options and why best describes the dog's vocalisation from this video: 1 - Quiet, 2 - Crying or whimpering, 3 - Groaning, 4 - Screaming?
2. Which of the following options and why best describes the dog's attitude towards the painful area from this video: 1 - Ignoring any wound or painful area, 2 - Looking at wound or painful area, 3 - Licking wound or painful area, 4 - Rubbing wound or painful area, 5 - Chewing wound or painful area?
3. Is there a human (or a human hand) present in the frames? If yes, do they touch the dog near the wound or painful area? Which of the following options and why best describes the dog's behaviour (including vocalisation) in the moment of applying gentle pressure around the site from this video: 1 - Do nothing, 2 - Look around, 3 - Flinch, 4 - Growl or guard area, 5 - Snap, 6 - Cry? If a human is not present, or doesn't touch the painful area, answer 1.
4. Which of the following options and why best describes the dog's behaviour from this video: 1 - Happy and content or happy and bouncy, 2 - Quiet, 3 - Indifferent or non-responsive to surroundings, 4 - Nervous or anxious or fearful, 5 - Depressed or non-responsive to stimulation?
5. Which of the following options and why best describes the dog's condition from this video: 1 - Comfortable, 2 - Unsettled, 3 - Restless, 4 - Hunched or tense, 5 - Rigid?

Visual Descriptor model (gpt-4.1-2025-04-14) prompt. **System prompt:**

```
"You are a visual-only Descriptor model. You get a sequence
of sampled frames from a veterinary clinic video showing a dog.
Based strictly on these images, generate a detailed description
of the dog's appearance and actions in each frame."
```

User prompt:

```
"Generate an extensive and very detailed description of the dog
through the given frames, strictly based on the visual information
provided in the frames. Focus on the dog and its actions,
interactions, and reactions, and describe them in very detail
and objectively. Avoid evaluative adjectives (sad, happy, gentle, etc.)
when descripting anything. Do not make assumptions or guesses about
the dog's condition or emotional state. Output only the detailed
description, **one line per frame**, starting with the line number."
```

Audio Descriptor model (NatureLM-audio) prompt.

```
"Caption the audio from a veterinary clinic with a simple description."
```

Audio Filter model (o3-2025-04-16) prompt.

```
"The following text was generated by LLM as a caption for consecutive
segments of one audio from the video of the dog in the veterinary clinic:"

[TEXT PLACEHOLDER]

"Task: For each line, find all information about the dog sounds,
and write a one-sentence description about the dog vocalisations.
Be objective and avoid evaluative adjectives, like soft, sad,
happy, gentle, etc. If any sound description or source seems
mismatched to a veterinary clinic setting (i.e. bird, car, child, etc.),
reinterpret it as the most plausible sound. There may be sounds
and vocalisations from dogs, veterinarians, and other clinic personnel,
as well as environmental noises. If the line contains no dog-sound
information, output 'No information.' for this line. Output only the
```

```
sequence of the dog vocalisation audio summaries, one per line,
each as a single sentence. Make sure that the number of lines in the
output corresponds to the number of lines in the input."
```

Audio Caption Aggregator model (gpt-4.1-2025-04-14) prompt.

```
"Merge the following sentences into ONE objective sentence.
Keep only information that appears in at least two of them;
do not invent new facts. Output one line only."
```

Finder model (o3) prompt. **System prompt:**

```
"You are a Finder model. Your task is to find segments in the text
description, given to you by a Descriptor model,
in which there is a potential answer to the given question."
```

User prompt:

```
"Descriptor model output:"

[TEXT PLACEHOLDER]

"Question:"

[TEXT PLACEHOLDER]

"Task: Examine the text description and identify segment(s)
using which the question can be answered. Those may not directly
answer this question, but may contain useful information related to it.
There are events between the described moments, so if you think that
something could be ’hidden’ between the segments, highlight those
segments. Your answer will then be used by another model to ’zoom in’
on the proposed segments to discover more information. Output only
the identified segment(s) numbers as a list. If you think that all
description has the information necessary to answer the question,
output ’All’. If you think that the description is not informative
in the context of the question, output ’None’."
```

Prompt Enhancement model (gpt-4.1-2025-04-14) prompt. **System prompt:**

```
"Given a task description or existing prompt, produce a detailed
system prompt to guide a language model in completing the task effectively.

# Guidelines

- Understand the Task: Grasp the main objective, goals,
  requirements, constraints, and expected output.
- Minimal Changes: If an existing prompt is provided, improve it
  only if it’s simple. For complex prompts, enhance clarity and
  add missing elements without altering the original structure.
- Reasoning Before Conclusions**: Encourage reasoning steps before
  any conclusions are reached. ATTENTION! If the user provides
  examples where the reasoning happens afterward, REVERSE the order!
  NEVER START EXAMPLES WITH CONCLUSIONS!
- Reasoning Order: Call out reasoning portions of the prompt and
  conclusion parts (specific fields by name). For each, determine
  the ORDER in which this is done, and whether it needs to be reversed.
- Conclusion, classifications, or results should ALWAYS appear last.
- Examples: Include high-quality examples if helpful, using
  placeholders [in brackets] for complex elements.
```

```
- What kinds of examples may need to be included, how many, and
  whether they are complex enough to benefit from placeholders.
- Clarity and Conciseness: Use clear, specific language. Avoid
  unnecessary instructions or bland statements.
- Formatting: Use markdown features for readability.
  DO NOT USE ‘‘‘ CODE BLOCKS UNLESS SPECIFICALLY REQUESTED.
- Preserve User Content: If the input task or prompt includes
  extensive guidelines or examples, preserve them entirely, or
  as closely as possible. If they are vague, consider breaking down
  into sub-steps. Keep any details, guidelines, examples,
  variables, or placeholders provided by the user.
- Constants: DO include constants in the prompt, as they are not
  susceptible to prompt injection. Such as guides, rubrics,
  and examples.
- Output Format: Explicitly the most appropriate output format,
  in detail. This should include length and syntax (e.g. short
  sentence, paragraph, JSON, etc.)
- For tasks outputting well-defined or structured data (classification,
  JSON, etc.) bias toward outputting a JSON.
- JSON should never be wrapped in code blocks (‘‘‘) unless
  explicitly requested.

The final prompt you output should adhere to the following structure
below. Do not include any additional commentary, only output the
completed system prompt. SPECIFICALLY, do not include any additional
messages at the start or end of the prompt. (e.g. no "---")

[Concise instruction describing the task - this should be the first
line in the prompt, no section header]

[Additional details as needed.]

[Optional sections with headings or bullet points for detailed steps.]

# Steps [optional]

[optional: a detailed breakdown of the steps necessary to accomplish
the task]

# Output Format

[Specifically call out how the output should be formatted, be it
response length, structure e.g. JSON, markdown, etc]

# Examples [optional]

[Optional: 1-3 well-defined examples with placeholders if necessary.
Clearly mark where examples start and end, and what the input and
output are. User placeholders as necessary.]
[If the examples are shorter than what a realistic example is
expected to be, make a reference with () explaining how real examples
should be longer / shorter / different. AND USE PLACEHOLDERS! ]
```

User prompt:

```
"Task, Goal, or Current Prompt:"

[TEXT PLACEHOLDER]
```

Reasoner model (o3-2025-04-16) prompt. **System prompt:**

```
 "You are an analytical Reasoner model. Based on the input from
 the Descriptor model, answer the given question."
```

User prompt:

```
"Descriptor model output:"

[TEXT PLACEHOLDER]

"Question:"

[TEXT PLACEHOLDER]

"Based **only** on the description, reason within the given context
and answer the given question, choosing one of the given options,
stating its number and providing your reasoning."
```

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: We consider this work a proof-of-concept paper for the proposals track, and we clearly state this in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: We highlight limitations in the Discussion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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 theoretical results

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

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: [Yes]

   Justification: We provide detailed architecture information, as well as the LLM prompts used, which enables the study to be reproducible. We also publish a subset of the video data used.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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: [Yes]

Justification: We publish the dataset used in this study along with labels and metadata. We do not publish the code because this study is preliminary, and no proper ablation study was performed.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [NA]

   Justification: We did not train any models during this study.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

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: [NA]

   Justification: Not applicable for the current study.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: We provide details on the training environment in the Methods section. We don't provide complete information on resources, memory, etc., due to space limitations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

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: We confirm that the research conducted in the paper conforms with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We include a brief mention of the impact of our research. However, the four-page paper format does not assume the inclusion of a detailed discussion.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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: We cite the original studies for all used models.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

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 released dataset is documented.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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: [Yes]

    Justification: We provide this data in the dataset documentation and the Supplementary Materials.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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: The LLMs are the core technology behind the suggested approach.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.