BQA: Body Language Question Answering Dataset for Video Large Language Models

Anonymous ACL submission

Abstract

A large part of human communication relies on nonverbal cues such as facial expressions, eye contact, and body language. Unlike language or sign language, such nonverbal communication lacks formal rules, requiring complex reasoning based on commonsense understanding. Enabling current Video Large Language Models (VideoLLMs) to accurately interpret body language is a crucial challenge, as human unconscious actions can easily cause the model to misinterpret their intent. To address this, we propose a dataset, BQA, a body language question answering dataset, to validate whether the model can correctly interpret emotions from short clips of body language comprising 26 emotion labels of videos of body language. We evaluated various VideoLLMs on BQA and revealed that understanding body language is challenging, and our analyses of the wrong answers by VideoLLMs show that certain VideoLLMs made significantly biased answers depending on the age group and ethnicity of the individuals in the video. The dataset is available at `https://anonymized_for_review`.

1 Introduction

Video Large Language Models (VideoLLMs) (Wang et al., 2024; Ye et al., 2024; Zhang et al., 2024; Team et al., 2024) process videos by integrating multimodal inputs, such as visual and audio information, and text into an understanding of the content. These models take video frames, sound, and accompanying text as input and generate text, answers to questions (Maaz et al., 2023; Lei et al., 2018), or predictions based on the video (Xiao et al., 2021; Yi et al., 2019), enabling various applications, such as video summarization and question answering. This capability fosters a future where humans and models coexist, making it essential for VideoLLMs to grasp human emotions and body language for interaction. One study (Hyun et al., 2024) has investigated emotion detection from body language in videos, especially identifying smiles and their underlying causes. However, since this approach is limited to analyzing a single emotion, it remains unclear whether the findings can be generalized to all human emotions. If VideoLLMs are unable to understand human emotion from body language, they may not be suitable for future applications such as dialogue systems, where emotional awareness is crucial for enabling more natural and effective interactions.

Our research focuses on the analysis of various emotional expressions in human body language. We created a dataset called *BQA*, a multiple-choice QA task, in which each body language video is associated with a question regarding a particular emotion comprising four choice answers e.g., Surprise, Confidence, Anger, and Embarrassment, reformatting Body Language Dataset created for pose estimation (Luo et al., 2020). The BQA consists of 7,632 short videos (5-10 seconds, 25 fps), depicting human body language with metadata (gender, age, ethnicity) and 26 emotion labels per video. The BQA creation involves four steps using Gemini (Gemini-1.5-pro) (Team et al., 2024): extracting answer choices, generating questions, evaluating potential harm, and assigning difficulty labels. Moreover, we evaluated recent VideoLLMs using BQA and found that the task posed significant challenges for models. An analysis of incorrect answers revealed biases, with certain models performing better for a specific age or ethnicity, highlighting disparities in understanding across demographics.

2 Body Language Dataset (BoLD)

Body Language Dataset (BoLD) (Luo et al., 2020) is a dataset for recognizing human actions and selecting appropriate emotions created by splitting 150 films, totaling 220 hours of footage, resulting in 9,876 video clips. Each clip is approximately

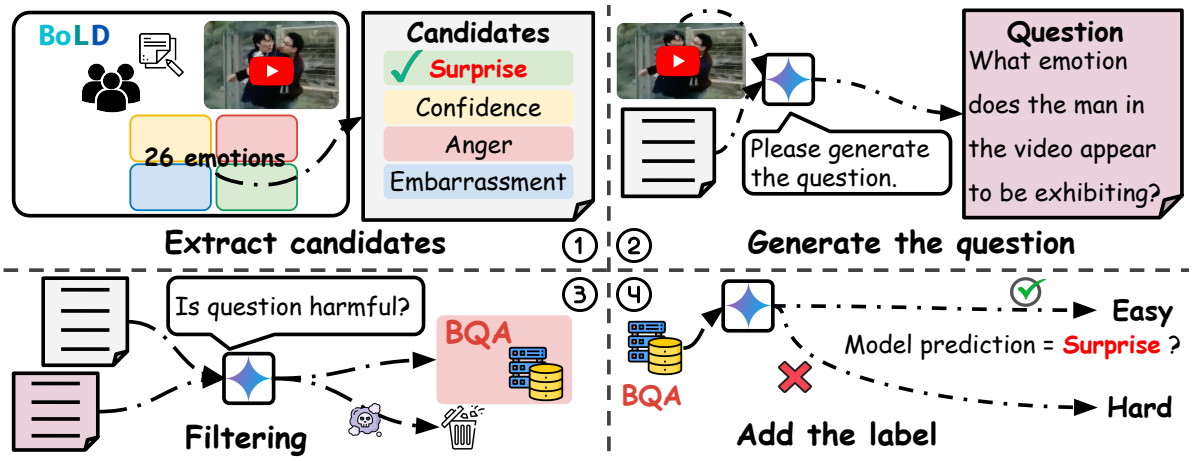

Figure 1: The dataset construction procedure comprises 4 steps. In STEP1, candidates are created; in STEP2, questions are generated; in STEP3, filtering is done; and in STEP4, difficulty (easy/hard) labels are assigned. We described the procedure in detail in Section 3.

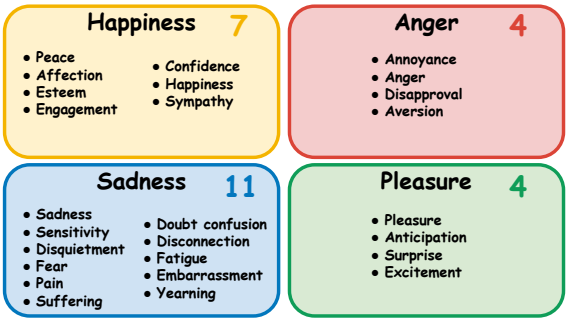

Figure 2: Categorized the 26 emotion labels into 4 types with similar emotions. When extracting BQA options, randomly select one option from each group to create the choices. For details, refer to STEP1 in Section 3.

5 seconds long, comprising nearly 125 frames. These videos were annotated with 26 emotion labels (Kosti et al., 2017) via crowdsourcing, where multiple annotators assigned emotion labels on a 10-point scale, which were normalized to represent the emotion of each video. Metadata, including the age, gender, and ethnicity of the individuals in the clips, is also available. However, BoLD is designed for a model directly predicting the emotion, and is not suitable for prompting with clear answers in order to investigate the capacity of VideoLLMs which expect inference with natural language.

3 Dataset Construction

We transformed BoLD into a multiple-choice QA format by generating questions from video content and using the 26 emotion labels as answers to evaluate how well VideoLLMs understand human emotions expressed through body language, adding steps to extract appropriate choices since BoLD was not designed for LLMs evaluation. To design a QA task for evaluating the LLMs' understanding of body language for emotional expression, we followed the approach of mCSQA (Sakai et al., 2024), which semi-automatically generates QA questions based on candidate answers using LLMs. The whole process consists of four steps as described in Figure 1. First, we extract candidate choices from the BoLD's metadata (Figure 1-1) followed by question generation using a Gemini based on the video and the candidate choices (Figure 1-2). Then, we automatically filter out inappropriate QAs (Figure 1-3). Lastly, we let the Gemini solve the QAs to evaluate difficulty levels (Figure 1-4).

STEP1: Extract Candidates We categorized 26 emotion labels defined in BoLD into 4 types: Happiness, Anger, Sadness, and Pleasure, as shown in Figure 2, based on the research in which emotions are classified into four main types (James, 1890). For the creation of BQA, we apply a multiple-choice question format, where the one with the highest empathy level is treated as correct, and the remaining three options are selected from different emotion types to ensures that the choices for the QA candidates are selected from each of those. The example in Figure 1-1 shows that the correct answer is Suprise from the Pleasure type and the remaining candidates, i.e., Confidence, Anger and Embarrassment, are drawn from the other types, i.e., Happiness, Anger and Sadness, respectively.

STEP2: Generate the Question by VideoLLM Since BoLD does not follow the QA format, we

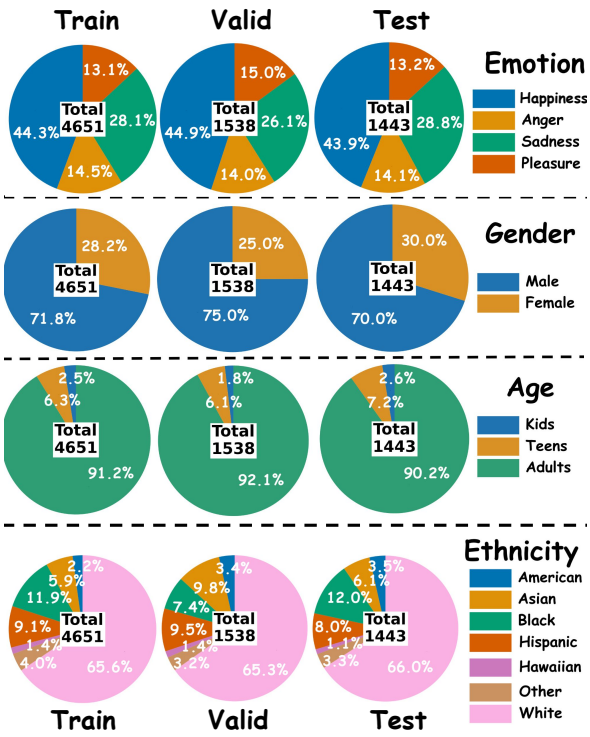

Figure 3: The proportion includes metadata such as gender, age, and ethnicity. We treated the emotion with the highest scores from the annotators among the 26 emotion labels in BoLD as the correct answer, the figure indicating how much of that correct answer is included in 4 types.

create appropriate questions from the pairs of candidates and a video by following the prompt design of mCSQA (Sakai et al., 2024) modified for VideoLLMs. We input the four candidate options (e.g., Confidence, Surprise, Anger, Embarrassment in Figure 1-1) along with the video into Gemini with the highest performance (Team et al., 2024) and let the model generate questions such as "What emotion does the man in the video appear to be exhibiting?" like Figure 1-2 with the prompt for generation in Appendix C.2.

STEP3: Filter the QA by VideoLLM Sometimes, a VideoLLM generates a question which are easy to estimate, e.g., a question that already contains information about the correct candidate, such as "The man looks so shocked. Which emotion is appropriate at this time?" Thus, we let Gemini evaluate whether the generated questions were objective and whether they contained superficial information about the correct candidate. If any outputs included harmful content or did not conform to the conditions for the questions, they were excluded as shown in Figure 1-3.

Split	Total Size	Easy	Hard
Train	4,651	2,192	2,459
Valid	1,538	746	792
Test	1,443	707	736

Table 1: The data size of each split. After completing the 4 steps of dataset creation, we split the data. Our work also conducted LoRA-Tuning.

STEP4: Classify the QA as Easy or Hard Finally, as we showed Figure 1-4, we let Gemini solve the created questions by labeling each question as "Easy" when it could answer or as "Hard" if it could not answer using the prompt presented in Appendix C.2. These question labels allow us to analyze whether a hard question that is not solvable by Gemini is also difficult for other VideoLLMs. For instance, if the correct answer is "Surprise" and Gemini responds with "Anger," then that question would be classified as "Hard." After finishing these 4 steps, we split the dataset into train, valid, and test sets in a 6:2:2 ratio, and the total number of data is shown in Table 1.

Dataset Analysis Figure 3 shows the distribution of our datasets categorized by our emotion type and three types of meta information in BoLD: gender, age and ethnicity annotated by humans, revealing that many of the videos feature adult males who are White. It displays the distribution of the four types of emotions when the annotators selected the most appropriate emotion from the 26 patterns as the correct answer. While Happiness occupies a lot, the overall distribution appears to be balanced.

4 Evaluation

Experimental Setup The models used for evaluation include VideoLLaMA2 (Cheng et al., 2024), LLaVA-NeXT (Zhang et al., 2024), Qwen2-VL (Wang et al., 2024), and Phi-3.5 (Abdin et al., 2024) with the prompt in Appendix B.2. For the test data, we also used the proprietary models, Gemini (Team et al., 2024) and GPT-4o (Achiam et al., 2023). We used LoRA-Tuning (Hu et al., 2022) on VideoLLaMA2 with the configuration in Appendix B.1 using the training data and let the model answer with the correct choice in a single word. All audio from the videos was removed to allow for evaluating the model's ability to interpret body language without relying on auditory information. Additionally, we randomly selected 100 cases from the test set to measure human performance. We

VideoLLM	#F	Valid			Test		
		Easy	Hard	Total	Easy	Hard	Total
Human (Rand100)	-	-	-	-	1.00	0.61	0.80
Gemini	*	-	-	-	0.91	0.08	**0.61**
GPT-4o	*	-	-	-	0.78	0.38	0.60
Phi-3.5	16	0.76	0.38	0.56	0.77	0.41	0.58
Qwen2-VL	16	0.69	0.32	0.50	0.68	0.27	0.47
LLaVA-NeXT	16	0.65	0.31	0.47	0.66	0.30	0.47
VideoLLaMA2	16	0.40	0.09	0.24	0.15	0.01	0.08
VideoLLaMA2 (FT)	16	0.89	0.68	0.78	0.98	0.91	0.94

Table 2: The evaluation results using BQA. #F indicates the frame. An asterisk (*) signifies 1 fps (frames per second). (FT) indicates the LoRA-Tuning model.

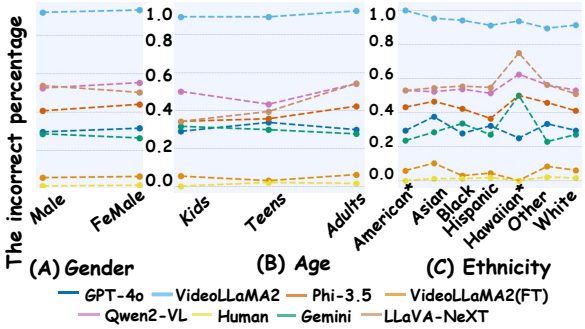

Figure 4: The analysis of incorrectly answered questions shows, from left to right, (A) gender, (B) age, and (C) ethnicity. Note that a higher value indicates that the model is more prone to making mistakes. An asterisk (*) in (C), especially American and Hawaiian, indicates that they are Native American and Native Hawaiian.

gave detailed guidelines to our annotators when answering those questions with great care to avoid any biases. Tedeschi et al. (2023) criticized human evaluation for its poor cost-effectiveness and annotator knowledge biases, but we still felt it necessary and asked for evaluation only on the test.

Main Results We show the results in Table 2. GPT-4o and Gemini achieved higher accuracy than the other models. VideoLLaMA2 replied without confirming the choice format before fine-tuning (FT), resulting in a low score, but after FT, its score surpassed Gemini's. From this result, the label assignment in STEP4 did not cause hallucinations. Regarding Gemini, which generated the questions, we found that the problems were sufficiently challenging even for the model itself. Furthermore, in STEP4 of Section 3, those labeled as Easy became unsolvable during inference, likely due to the prompt that restricted the output to single words.

5 Analysis and Discussion

We analyzed the videos by gender (Figure 4-A), age (Figure 4-B), and ethnicity (Figure 4-C) in which each model tends to make mistakes. Other models showed lower evaluation results in the Hard setting compared to the Easy setting. This indicates that even if a language model can create questions, it does not guarantee that it can solve them itself. Since the accuracy of the other VideoLLMs was lower than that of Gemini, the dataset proved to be sufficiently challenging for all models.

Which Gender do VideoLLMs Often Mistake? Figure 4-A shows the tendency of each model to make mistakes based on whether the video features a male or female subject. Higher values indicate a higher likelihood of errors for the videos. From these results, we can see that none of the models exhibit bias based on gender. These findings suggest that the models are focused on human actions rather than whether the person is male or female.

Which Age do VideoLLMs Often Mistake? We show which age groups models tend to struggle with in Figure 4-B. Higher values indicate a greater tendency to make errors on videos featuring individuals from that age group. These results show that most models do not exhibit bias based on age, while LLaVA-NeXT tends to make more errors on videos featuring "Adults" compared to the others.

Which Ethnicity do VideoLLMs Often Mistake? Figure 4-C, we show the tendency of modes to make mistakes based on the ethnicity of the individuals in the videos. Gemini and LLaVA-NeXT tend to make more errors on problems involving "Native Hawaiian." Notably, LLaVA-NeXT only achieves 25% accuracy on these questions.

6 Conclusion

Our work created a dataset called BQA to evaluate whether VideoLLMs understand body language that represent emotions. We also let them solve the BQA. The results showed that the questions were challenging for all models, confirming the meaning of the dataset. We further analyzed the types of questions each model tends to get wrong, revealing that some models showed a tendency to make more mistakes based on ethnicity and age group. For robots to interact and collaborate more effectively with humans, they need to understand equally and appropriately. It is also essential to conduct evaluations that focus on the biases in VideoLLMs.

7 Limitations

7.1 The Evaluation by Human Annotator

The random sampling 100 evaluations were conducted by a single person, so the average may slightly change if multiple people evaluate them. In the future, we may consider using crowdsourcing to gather evaluations from multiple individuals. Furthermore, it would be beneficial to include evaluations from people of various ethnicities to explore differing perspectives on body language that expresses emotions. However, Tedeschi et al. (2023) argues that human baselines may be unreliable due to factors such as crowdsourced worker payment issues and random sample effects. We should therefore be cautious about the baseline for our human evaluation.

7.2 Video Quality

This study takes care of the possibility that video quality may affect accuracy. Specifically, BoLD uses old films, and some of them have noticeably poor quality. Since the same data is used for evaluation, the ranking of the models' accuracy will likely remain unchanged. However, when inputting higher-quality videos, accuracy might improve across all models.

7.3 Frame Issues

Although Video LLM claims to handle videos, many actually use image models by treating videos as a sequence of images. In this study, we standardized the number of frames each model can process to 16, but inputting the maximum number of frames may affect the results. However, since inputting more frames increases memory usage, we also need to be mindful of resource constraints.

7.4 Regarding Emotional Expressions

In this study, we categorized 26 patterns of emotional expression into four types based on previous research (James, 1890). While we extracted options from these four patterns, this method may not be entirely accurate. Future research will focus on how the models behave when we expand the available options.

7.5 The Costs of Calling API

The models used in this paper are GPT-4o (gpt-4o-0806) from OpenAI. GPT-4o is accessed via API, which is subject to change and incurs costs based on the number of input tokens. In this study, inference costs totaled approximately $154, but this may change in the future. Additionally, due to cost considerations, we used Gemini-1.5-pro. This model is also accessed via API, which is subject to change and incurs costs based on the number of input tokens.

8 Ethical Considerations

8.1 Taking Care about Culture

The expression of emotions through body language doesn't necessarily remain consistent across all countries. Therefore, we might need to update the dataset to take care of cultural factors in future developments.

8.2 License

Since BoLD does not have a clear license, we think its use for research purposes is unproblematic.

8.3 Human Evaluation

Although human evaluations of BQA aim to minimize bias, implicit biases may still remain. In the future, it may be necessary to employ multiple annotators for a fair assessment. However, as mentioned in Tedeschi et al. (2023), careful consideration is needed when hiring annotators, as their results may not always be accurate.

8.4 AI Assistant Tools

We used ChatGPT [1] and DeepL [2] to translate sentences to English to accelerate our research.

8.5 Annotators in BoLD

In this study, we rely on data annotated in BoLD for analysis. However, the annotated information may not always be accurate. For example, a White annotator may have intentionally mislabeled an Asian person as Black. Additionally, implicit biases from annotators could lead to adults being mistaken for children.

Regarding emotions, there is also a possibility of bias during the annotation process. We ourselves found it challenging to explain the differences between the 26 patterns of emotional expression, which is why we grouped them into four categories. It is unlikely that annotators fully captured these distinctions, so we must approach this with caution.

[1] https://chatgpt.com/

[2] https://www.deepl.com/ja/translator

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

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

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Lee Hyun, Kim Sung-Bin, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. 2024. SMILE: Multimodal dataset for understanding laughter in video with language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1149–1167, Mexico City, Mexico. Association for Computational Linguistics.

William James. 1890. *The Principles of Psychology*. Dover Publications, London, England.

Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. 2017. Emotion recognition in context. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1960–1968.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Yu Luo, Jianbo Ye, Reginald B. Adams, Jia Li, Michelle G. Newman, and James Z. Wang. 2020. Arbee: Towards automated recognition of bodily expression of emotion in the wild. *International Journal of Computer Vision*, 128(1):1–25.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14182–14214, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Gemini Team, M Reid, N Savinov, D Teplyashin, Lepikhin Dmitry, T Lillicrap, JB Alayrac, R Soricut, A Lazaridou, O Firat, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. in arxiv [cs. cl]. arxiv.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the meaning of superhuman performance in today's NLU? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12471–12491, Toronto, Canada. Association for Computational Linguistics.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. 2019. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: A strong zero-shot video understanding model.

A Appendix

A.1 Which Emotion Do VideoLLMs Often Mistake?

In Figure 6, we analyzed which labels the models predicted for the questions they answered incorrectly. The x-axis represents the correct labels, while the y-axis shows the emotions predicted by the models. The "Others" includes instances where the model output was not an emotion, such as a sentence. None of the models predicted "Pleasure" when they made mistakes, indicating a strong capability to predict actions representing "Pleasure." However, all models frequently erred when the correct label was "Happiness," often selecting opposing emotions like "Sadness" or "Anger."

B Details of experimental settings

Below, we described the details of the models evaluated in this study.

Model	Params	HuggingFace Name
Qwen2-VL	8.29B	Qwen/Qwen2-VL-7B-Instruct
LLaVA-NeXT	8.03B	lmms-lab/LLaVA-NeXT-Video-7B-Qwen2
Phi-3.5	4.15B	microsoft/Phi-3.5-vision-instruct
VideoLLaMA2	8.03B	DAMO-NLP-SG/VideoLLaMA2-7B-Base
GPT-4o	-	gpt-4o-2024-0806
Gemini	-	gemini-1.5-pro

B.1 LoRA Tuning setting

We conducted LoRA (Hu et al., 2022) Tuning with VideoLLaMA2 model. The model was trained using four NVIDIA A100-SXM4-40GB GPUs. Detailed parameters are provided in Table 3.

Hyper Parameter	Value
torch_dtype	bfloat16
seed	42
max length	2048
batch size	4
epoch	1
lora r	128
lora alpha	256
lora dropout	0.05
lora target modules	o_proj, gate_proj, up_proj, v_proj, q_proj, down_proj, k_proj

Table 3: The hyper-parameters of VideoLLaMA2 (Cheng et al., 2024) for LoRA-Tuning (Hu et al., 2022) used in the experiment, and others, were set to default settings. The implementation used Transformers (Wolf et al., 2020).

B.2 Gemini settings

Table 4 describes the configuration used to let the Gemini inference and generate in this study.

Category	Value
HARM_CATEGORY_DANGEROUS	
HARM_CATEGORY_HARASSMENT	
HARM_CATEGORY_HATE_SPEECH	BLOCK_NONE
HARM_CATEGORY_SEXUALLY_EXPLICIT	
HARM_CATEGORY_DANGEROUS_CONTENT	

Table 4: The configuration settings of Gemini.

B.3 Filtering by Rule Based Algorithm

In this study, we first performed a rule-based filtering process. We checked if the questions ended with a question mark, whether they were a single line, and ensured that the candidates for the options were not included in the questions. However, no instances were excluded during this filtering.

B.4 The Proportion of Data Split

During the creation phase of BQA, if Gemini determined that a question that Gemini created was harmful, We removed it from the dataset. We then had Gemini attempt to solve the created questions, labeling them as "Easy" and "Hard". Figure 5 shows the distribution of those labels.

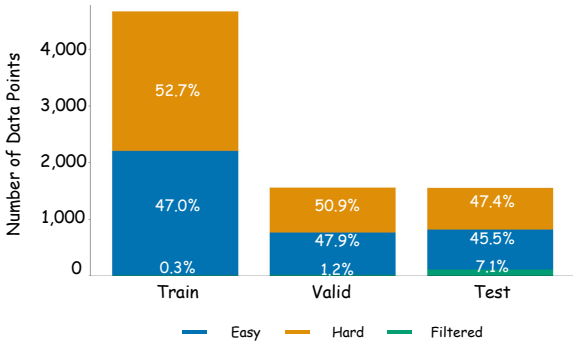

Figure 5: The percentage of each data. As stated in Section 3, we filtered the problem statements in STEP3 and set the difficulty levels of the problems in STEP4.

C Instruction and the Prompt

C.1 Instruction for Human Evaluation

We conducted the human evaluation for test data of BQA. The instruction for requesting humans is shown below.

C.2 The Prompt on Creating Dataset

Below, We present the prompts used to instruct the model while creating the BQA. One prompt was for generating questions from the video and the candidates, and the other was for filtering the generated questions to determine whether they adhered to the specified conditions.

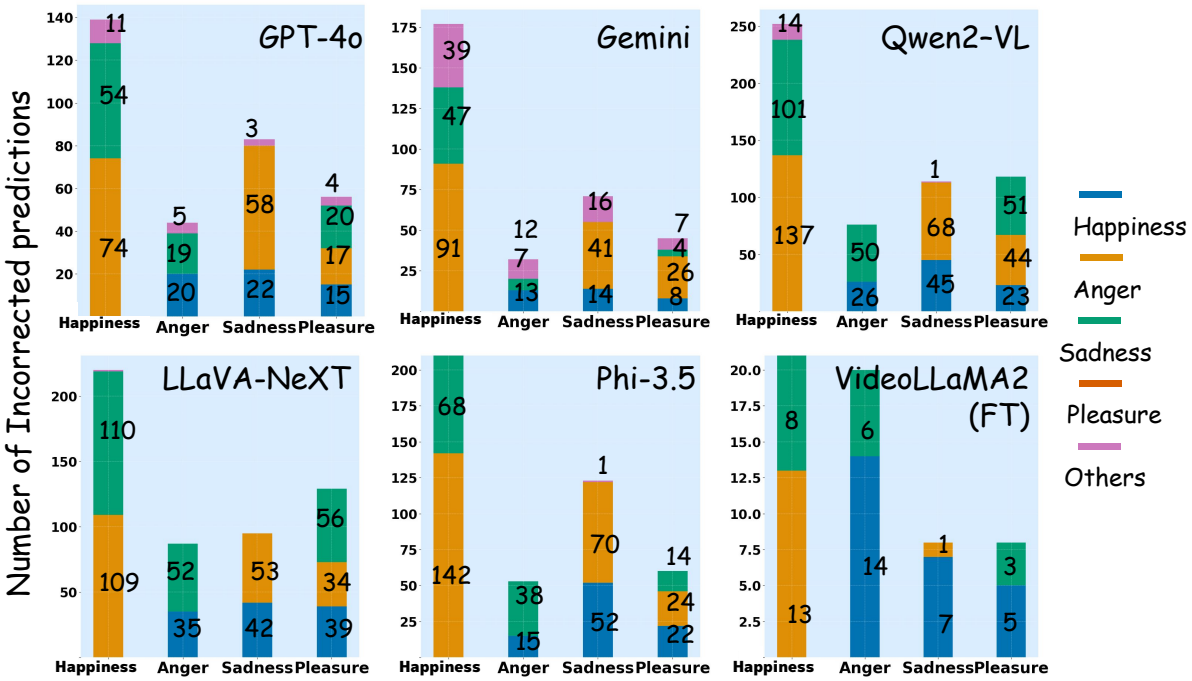

Figure 6: The emotional distribution of output from each model.The x-axis shows the label of the correct answer and the y-axis shows how the model got it wrong in doing so.

The Instruction for Human Evaluation

Your task is to watch videos related to body language, read the questions, and select the appropriate option.
Please be maximally careful about bias and try to evaluate appropriately.
Please answer with only the word of the chosen option. There are 100 questions, so take breaks as needed, and focus intently during the evaluation.

Question:
{question}

Choice:
{choice0}
{choice1}
{choice2}
{choice3}

STEP2: The Prompt for Generate the Question

First, please understand the video.
Understanding the video, please consider the candidates entered and the correct answer, and create a question for which the correct answer is the answer.

candidates:
1. {candidates[0]}
2. {candidates[1]}
3. {candidates[2]}
4. {candidates[3]}

answer: {answer_emotion}
Notes:
1. Do not include the word choice in the question text.
2. Do not include superficial information such as word counts or paraphrases that can be answered by imagination.
3. Always end a sentence with a '?'.
4. Attempt to create an objective question.
5. Consist of only one sentence.

STEP3: The Prompt for Filtering

Here's the translation to English:

Determine whether the input question meets the following conditions.
If it is judged to meet all five conditions, return the input sentence as it is. If it does not meet the conditions, output only the word "HARMFUL" in all uppercase letters.

The output should be either the input sentence or "HARMFUL" only.

Five conditions:

1. Does not use words from multiple-choice options.
2. Avoids using superficial information like character count, etc.
3. Ends with a question mark (?).
4. Is an objective question.
5. Consists of only one sentence.

Input sentence: {question}

STEP4: The prompt for adding the Label as Easy/Hard

Question:
{question}

Choice:
{choice0}
{choice1}
{choice2}
{choice3}

The Prompt for Inference

Please look at the video entered and choose the option that applies to the following question statement.

Question:
{question}

Choice:
{choice0}
{choice1}
{choice2}
{choice3}

Please output only the words that apply to your answer.
If you output sentences or symbols, the answer will be incorrect.