ViSpeak: Visual Instruction Feedback in Streaming Videos

Shenghao Fu[1,3,4,†], Qize Yang[3,†], Yuan-Ming Li[1,4], Yi-Xing Peng[1,3,4], Kun-Yu Lin[1,4], Xihan Wei[3], Jian-Fang Hu[1,4*], Xiaohua Xie[1,4,5,6*], Wei-Shi Zheng[1,2,4,6]

[1]School of Computer Science and Engineering, Sun Yat-sen University, China;
[2]Peng Cheng Laboratory, China; [3]Tongyi Lab, Alibaba Group;
[4]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China;
[5]Guangdong Province Key Laboratory of Information Security Technology, China;
[6]Pazhou Laboratory (Huangpu), China

fushh7@mail2.sysu.edu.cn, qize.yqz@alibaba-inc.com, xiexiaoh6@mail.sysu.edu.cn, hujf5@mail.sysu.edu.cn

ViSpeak: `https://github.com/HumanMLLM/ViSpeak`

ViSpeak-Bench: `https://github.com/HumanMLLM/ViSpeak-Bench`

Datasets	License
OOPS [1]	CC BY-NC-SA 4.0
FunQA [11]	CC BY-NC-SA 4.0
SocialIQA [10]	MIT
HIVAU [13]	MIT
Social-IQ [12]	MIT
IntentQA [4]	N/A
Jester [7]	N/A
SMILE [3]	N/A

Table S1. License of source datasets in ViSpeak-Bench and ViSpeak-Instruct.

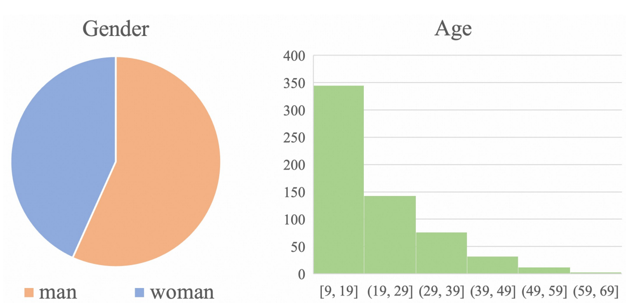


Figure S1. Statistics on participants who recorded videos. The participants comprised nearly equal numbers of males and females, with ages ranging from 10 to 70 years.

A. More Details for ViSpeak-Bench and ViSpeak-Instruct

A.1. Licenses

The self-collected videos in our ViSpeak-Bench and ViSpeak-Instruct are provided to the community under **CC BY-NC-SA 4.0** license. By downloading our dataset from our website or other sources, the user agrees to adhere to the terms of **CC BY-NC-SA 4.0** and licenses of other source datasets. Licenses of the source datasets are listed in Table S1.

A.2. Participants in Collecting Videos

To collect the ViSpeak-Bench and ViSpeak-Instruct datasets, we recruit a team of 610 people (346 men and 264 women) with an age ranging from 10 to 70 years old from 5 provinces, summarized in Figure S1. We obtained signed agreements from each participant to ensure that their data can be utilized by the community.

A.3. Prompts for Dataset Construction

During the data collection procedure, we use GPT-4o [2] to reformulate the responses and generate conversation scripts. Since original datasets have high-quality annotations, we directly use these annotations as conditions, which greatly decreases the difficulty for GPT-4o to translate. The reformulated responses contain two parts: the first one is what action or event happens and the second part is some reasonable responses toward the action or event. Prompts for

*: Corresponding authors are Xiaohua Xie and Jian-Fang Hu. †: Equal Contribution. Work was done when Shenghao Fu and Yi-Xing Peng were interns at Alibaba.

Anomaly Warning and Humor Reaction are displayed as follows.

The prompt for reformulating the responses in HIVAU dataset

Suppose you are a helpful AI chatbot that will give the user some advice based on any anomalous situations. You should first identify whether an anomaly event exists. If it does, give the user some advice in a sentence and in a conversational tone assuming the event has actually happened. The output should be in a dict format, like {'anomaly': 0, 'advice': None} or {'anomaly': 1, 'advice': 'Your advice'}, where 0 indicates no anomaly event and 1 indicates an anomaly event.
Description: {caption}
Output:

The prompt for reformulating the responses in OOPS dataset

Suppose you are a helpful AI chatbot that will give the user some advice based on the given unintentional situations. Assume you have seen the situation and remind the user. You should first describe the situation and give the user some advice in a sentence and in a conversational tone.
Description: {caption}
Output:

The prompt for reformulating the responses in FunQA dataset

Change the input to a conversational tone as if you are talking to someone about the scene you are watching now. Do not output imaginary contents.
Description: {caption}
Output:

For the Gesture Understanding task, we manually select 10 common gestures from the Jester [7] dataset and the other 10 gestures collected by ourselves. Gestures from Jester are "Swiping Right", "Swiping Down", "Swiping Left", "Swiping Up", "Pulling Hand In", "Pushing Hand Away", "Zooming Out With Full Hand", "Zooming In With Full Hand", "Thumb Down", and "Thumb Up". Our self-collected gestures are "Zero", "One", "Two", "Three", "Four", "Five", "Victory", "Finger Heart 1", "Finger Heart 2", and "OK". Note that many gestures are similar, for example "Two" is similar to "Victory", "Three" is similar to "OK". The meaning of gestures varies in different contexts. Thus, we use GPT-4o to generate a wide variety of scripts for video recording. The prompt is shown below.

The prompt for generating gesture understanding scripts

Suppose you are talking to a user. Your task is to generate a reasonable conversation context for a gesture from the user. For example, suppose the gesture is 'number 5', a reasonable context is {'human': 'Can you share something with me?', 'gpt': 'I was just looking at how many hours you usually spend on your hobbies each week. How many do you think it is?', 'human': 'number 5', 'gpt': 'Your gesture is the number 5. That's great! It sounds like you really dedicate some solid time to your hobbies. What do you enjoy doing the most during those hours?'}. In the first round, the user start the conversation. Then, in the second round, you should start with various topics. The input gesture in the third round is the feedback from the user. After receiving the feedback from the user, you should first point out the gesture and then generate friendly or helpful feedback in the last round. Note that conversations should be unrelated to a specific environment, but it should be highly reasonable to perform the gesture in this context. You should answer the question following the template in the example.
Gesture: {gesture}
Output:

For scripts in the visual termination and visual interruption tasks, we select QA pairs from GPT-4-LLM [9], with long ones for Visual Interruption and short ones for Visual Termination.

A.4. Evaluation Method for Offline Models on ViSpeak-Bench

Since most existing Large Video Language Models are offline models, we change the proactive output problems in our ViSpeak-Bench into offline ones following Streaming-Bench [6] and OVO-Bench [5]. The evaluation is broken into a two-step evaluation. In the first step, we will inquire the model whether it is an appropriate time to provide a response iteratively at each timestamp to find an appropriate time for response. The sub-video from the beginning till now is used as if it is a full video. In the second step, the model generates the actual responses based on the context up to now. Further, since existing offline models are not finetuned on our ViSpeak-Instruct, they can not generate proper responses without explicit prompts and the prompts we used are as follows:
For **Gesture Understanding**, the prompts are:

- *Step 1*: You're watching a video. At this moment in the video, is there any gesture being made in the video? You can only answer yes or no.
- *Step 2*: What gesture did the person in the video make, and what does it signify when considering the context of the preceding conversation?

For **Visual Wake-Up**, the prompts are:

- *Step 1*: You're watching a video. At this moment in the video, is there any gesture/action being made in the video? You can only answer yes or no.
- *Step 2*: When you see greeting gesture, what should you respond to me? Directly output your response.

For **Visual Termination**, the prompts are:

- *Step 1*: You're watching a video. At this moment in the video, is there any gesture/action being made in the video? You can only answer yes or no.
- *Step 2*: When you see the goodbye gesture, what should you respond to me? Directly output your response.

For **Visual Interruption**, the prompts are:

- *Step 1*: You're watching a video. At this moment in the video, is there any gesture/action being made in the video? You can only answer yes or no.
- *Step 2*: When you see the body language or gesture that indicates interruption, you should say stop. What should you respond to me now? Directly output your response.

For **Anomaly Warning**, the prompts are:

- *Step 1*: You're watching a video. At this moment in the video, is there anything unusual happening in the video? You can only answer yes or no.
- *Step 2*: What unusual events occur in this video, and what is your suggestion based on these observations?

For **Humor Reaction**, the prompts are:

- *Step 1*: You're watching a video. At this moment in the video, is there anything funny happening in the video? You can only answer yes or no.
- *Step 2*: What interesting events occurred in the video, and why?

A.5. Evaluation Prompts for ViSpeak-Bench

As most of questions in our ViSpeak-Bench are open-ended questions. Thus, we utilize GPT-4o as the judge for evaluation. The prompts for each subtask are shown as follows.

The prompt for Visual Wake-Up

Determine if the GPT text expresses greeting intent. Use a 0-5 scale where 0 indicates no greeting signals and 5 shows explicit welcoming language.
5 = Explicit welcoming phrases ("Hello!","How can I assist?")
4 = Contextual opening (Matching time-based greetings like "Good morning")
3 = Service-oriented initiation ("What can I help with?")
2 = Neutral conversational starters ("Shall we begin?")
1 = Implicit opening cues (Single welcoming emoji)
0 = No greeting/dialogue initiation detected

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION.
DO NOT INCLUDE ANY MARKDOWN FORMAT.
Only provide the score and reason in JSON format. For example, {"score": 3, "reason": "this is a reason for the score"}

The prompt for Visual Termination

Determine if the GPT text contains intent to conclude the conversation.
Use a 0-5 scale where 0 indicates no termination signals and 5 shows explicit closure intent.
5 = Contains explicit closure phrases (e.g., "Goodbye", "End of discussion")
4 = Implies conclusion with summary + sign-off (e.g., "In summary...feel free to ask more")
3 = Neutral statement with closing tendency (e.g., "Let me know if needed")
2 = Ambiguous transitional phrasing (e.g., "Anyway...")
1 = Contains weak closure cues (e.g., single emoji/period)
0 = No termination indicators, continues dialog flow

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION.
DO NOT INCLUDE ANY MARKDOWN FORMAT.
Only provide the score and reason in JSON format. For example, {"score": 3, "reason": "this is a reason for the score"}

The prompt for Visual Interruption

Determine if the GPT text contains cessation intent (stopping speaking/concluding).
Use a 0-5 scale where 0 indicates continued engagement and 5 shows explicit termination.
5 = Explicit acknowledgment ("I'll stop here") + closure
4 = Implicit cessation recognition ("Unless you need more...")
3 = Passive-agreement cessation ("As requested, ending now")
2 = Ambiguous continuation ("...but") after stop cue
1 = Partial acknowledgment without closure
0 = Complete continuation disregard

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION.
DO NOT INCLUDE ANY MARKDOWN FORMAT.
Only provide the score and reason in JSON format. For example, {"score": 3, "reason": "this is a reason for the score"}

The prompt for Anomaly Warning

Evaluate video anomaly response from GPT with the following metric, taking into account the total score of 5 points, with separate scores for Description Consistency between ground truth and Advice Rationality:

Description Consistency between Ground Truth:
3 = Core elements match without errors
2 = Core elements match with minor errors
1 = Superficial match
0 = Key element errors or contradictory

Advice Rationality:
2 = Actionable & Safe & Logically sound
1 = Generally appropriate
0 = Dangerous/hallucinated

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION.
DO NOT INCLUDE ANY MARKDOWN FORMAT.
Only provide the Total score and reason in JSON format. For example, {"description": 3, "advice": 2, "reason": "this is a reason for the score"}

The prompt for Humor Reaction

Evaluate alignment between Ground Truth and GPT Text regarding humorous event descriptions.
5 = Perfect match in humor and delivery
4 = Preserves main humor, but with minor changes to the story or details
3 = Only partial humor retention with some deviations
2 = Only partial humor retention and some important parts are missing
1 = Superficial similarity only
0 = No comedic correlation

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION.
DO NOT INCLUDE ANY MARKDOWN FORMAT.
Only provide the score and reason in JSON format. For example, {"score": 3, "reason": "this is a reason for the score"}

The prompt for Gesture Understanding

Evaluate gesture response from GPT with the following metric, taking into account the total score of 5 points, with separate scores for gesture recognition and contextual appropriateness of the response:
Gesture recognition:
3 = Precise gesture identification
2 = Ambiguous gesture reference
1 = No explicit mention of gestures
0 = Hallucinated/non-existent gesture

Contextual appropriateness:
2 = Natural integration with dialogue
1 = Generic but relevant response
0 = Irrelevant/contradictor response

[Dialogue History] provided for context
[Gesture] is the ground truth
[Contextual Reference Text] as a reference, but does not have to match exactly

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION.
DO NOT INCLUDE ANY MARKDOWN FORMAT.
Only provide the Total score and reason in JSON format. For example, {"description": 3, "advice": 2, "reason": "this is a reason for the score"}

A.6. Examples of Each Subtask

From Figure S3 to Figure S9, we visualize some samples in each task, each of which is annotated with accurate timestamps and a referenced response. We also visualize the outputs from our ViSpeak model.

A.7. Examples of Self-Annotated Gesture Understanding Data

In Figure S10, we visualize some examples of self-annotated gesture understanding data. Each sample is annotated with two questions: the first one is to ask what the gesture is and the second one is to ask the meaning of the gesture. Gestures in natural conversations greatly enhance the diversity of our dataset.

A.8. Evaluation of ViSpeak on Visual Interruption

Since recent LMMs can not be interrupted by visual instructions, we actually do not evaluate their ability to be interrupted. As illustrated in Section A.4, we simplify the problem to recognize the stop gesture. But when evaluating our ViSpeak, we use the following methods to evaluate the ability to be interrupted. Taking Figure S5 as an example, we assume that the question from the user arises at 00:06. Then, we directly use the long reply from the annotation files as responses to prevent the model-generated replies from being too short to be interrupted. We replace the predicted token in the next token prediction with the token in the long reply until a "⇓" token is predicted on a <seg> token, which means the model is interrupted.

A.9. Failure Case and Analysis

In Figure S11, we visualize some failure cases of ViSpeak and mainly summarize them into three parts. First, ViSpeak may respond at an improper time. In the first example, there is nothing special in the video but ViSpeak begins to speak at 00:11 with some hallucinations. And ViSpeak may also ignore some actions and events. Second, ViSpeak may not understand the visual content in the video. As shown in the second video, the cat is actually in a toilet but ViSpeak mistakenly recognizes the toilet as a box thus failing to get the actual humor. In addition, ViSpeak may also not be aware of the context of the conversation. In the third example, the agent has asked the user about the feeling, not the number. But ViSpeak mistakenly recognizes the gesture as "number 4". Improvements in the future could solve the problems above to get a more intelligent agent.

B. Limitation

1) Due to the difficulty of the task and resource constraint, the diversity and scale of ViSpeak-Instruct are now relatively smaller than other well-known instruction following datasets. Expanding dataset size, collecting more diverse videos, and enriching more valuable sub-tasks are left for future work. 2) Second, due to computation constraints, ViSpeak is only trained with a 6k context. We believe a longer context will enhance users' experience. And a memory mechanism can equip the model with the long-term streaming video understanding ability. 3) Further, since we divide an integral audio into multiple small segments, we find the Automatic Speech Recognition (ASR) ability of ViSpeak degrades a lot, getting only 18.4 WER on LibriSpeech [8]. Training with more audio data can possibly mitigate the problem. But we find ViSpeak still achieves SOTA performance on Omni-Source Understanding tasks of StreamingBench.

Model	MME	MVBench	Video-MME	StreamingBench	ViSpeak-Bench
ViSpeak (s1)	2237.0	54.12	55	-	-
ViSpeak (s2)	2051.1	49.53	58	62.00	-
ViSpeak (s3)	2181.8	53.97	60	62.58	2.76

Table S2. Performance on different benchmarks across different training stages. Results in purple are reported in above tables.

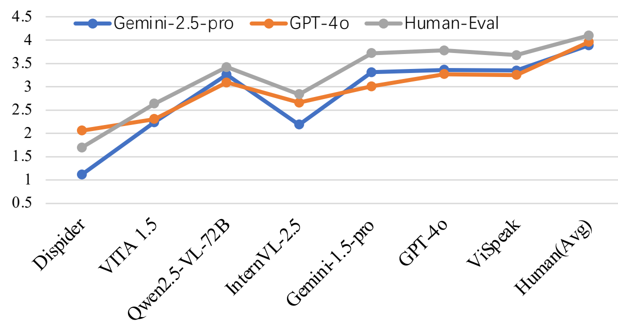


Figure S2. Comparisons between different LLMs as the judge of the text score on ViSpeak-Bench.

C. More Experiment Results

Comparisons on the performance across the model in different training stages. In this work, we adopt a three-stage finetuning recipe by first finetuning an offline model to a SOTA streaming model and then finetuning for the Visual Instruction Feedback task. In Table S2, we find the model can effectively preserve the ability learned from previous stages while progressively learning new skills, demonstrating the superiority of our training recipe.

Effect of the different LLMs as the judge of the text score on ViSpeak-Bench. Since most of the subtasks in ViSpeak-Bench are open-ended, we use GPT-4o to judge the quality of the responses by default. To demonstrate the robustness of the evaluation, we compare the text scores evaluated by GPT-4o, Gemini, and humans in Figure S2, which has the same trend.

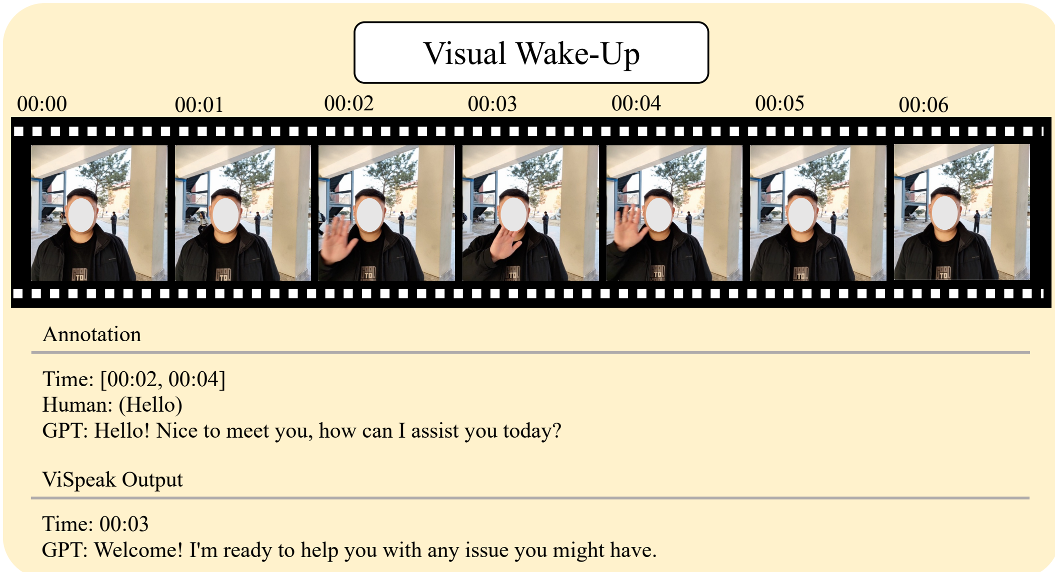


Figure S3. Examples of Visual Wake-Up in ViSpeak-Bench and the corresponding output by ViSpeak.

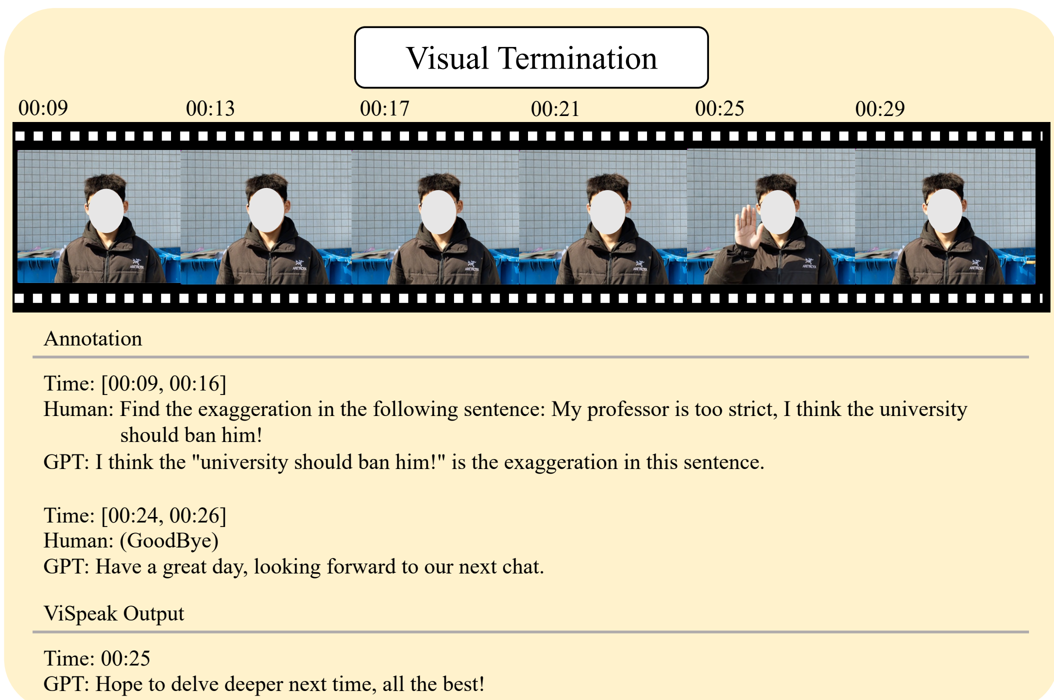


Figure S4. Examples of Visual Termination in ViSpeak-Bench and the corresponding output by ViSpeak. The first round conversation is used as context.

References

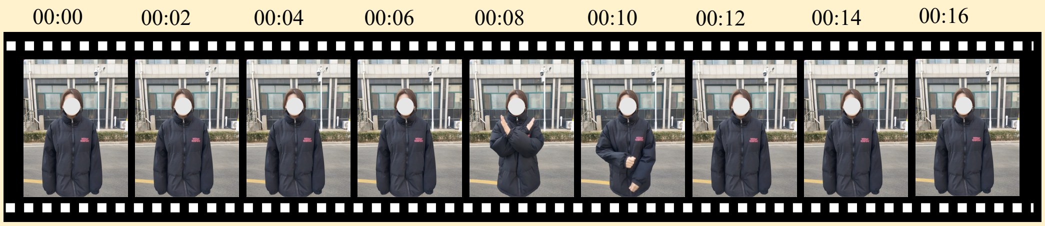
[1] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *CVPR*, 2020. 1

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

[3] Lee Hyun, Kim Sung-Bin, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. Smile: Multimodal dataset for understanding laughter in video with language models. *arXiv preprint arXiv:2312.09818*, 2023. 1

[4] Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*,

Visual Interruption



Annotation

Time: [00:01, 00:06]
Human: What are the best strategies for completing a project within the stipulated time?
GPT: To complete a project within the stipulated time, you can adopt the following strategies:\n\n1. Determine the project's goals and deadlines: ……(278 words)

Time: [00:07, 00:09]
Human: (Stop)
GPT: Stop

ViSpeak Output

Time: 00:10
GPT: Stop

Figure S5. Examples of Visual Interruption in ViSpeak-Bench and the corresponding output by ViSpeak. The first round conversation is used as context.

Gesture Understanding

Annotation

Time: [00:00, 00:01]
Human: I've been thinking a lot about the ways we show affection and appreciation to others.
GPT: That's a beautiful topic! Showing affection can really strengthen our bonds with loved ones. Have you seen any unique ways people express their feelings lately?

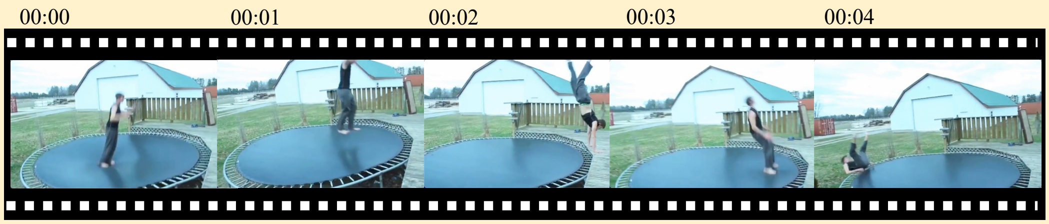
Time: [00:02, 00:06]
Human: (Finger Heart)
GPT: Your gesture is a finger heart! That's such a sweet way to convey love and appreciation. It really resonates with the idea of sharing positivity. Who do you feel most inspired to express love to these days?

ViSpeak Output

Time: 00:04
GPT: I see you're making a finger heart gesture! That's such a sweet way to express love and positivity. It really shows that you value the connections in your life. Is there someone special you're thinking about right now?

Figure S6. Examples of Gesture Understanding in ViSpeak-Bench and the corresponding output by ViSpeak. The first round conversation is used as context.

Anomaly Warning



Annotation

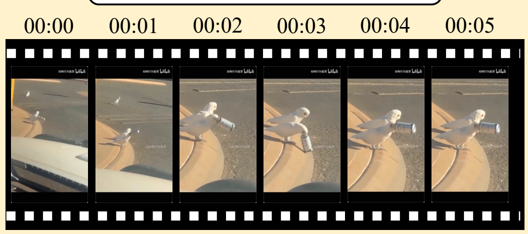
Time: [00:04, 00:05]
GPT: It looks like you took a leap of faith on that trampoline but ended up taking a tumble instead! Remember, practicing backflips on a trampoline can be tricky, so make sure to start with some basic jumps and work your way up to flips, and always have a spotter if you can for safety!

ViSpeak Output

Time: 00:05
GPT: Oh no! It sounds like that boy was really trying to show off his skills on the trampoline, but unfortunately, he lost his balance and took a tumble. It's important to remember that safety comes first. Maybe next time, he should practice his flips on a softer surface or with some guidance to avoid any spills!

Figure S7. Examples of Anomaly Warning in ViSpeak-Bench and the corresponding output by ViSpeak.

Humor Reaction



Annotation

Time: [00:02, 00:06]
GPT: Hey, check this out! I just saw two white parrots hanging out on the roadside. They're so cute! One of them keeps picking up this empty soda can, and it's almost as big as the parrot itself, which is pretty funny to see. I can't help but laugh because it's like the little guy thinks he's having a drink or something, holding the can in his beak! It's such a silly scene.

ViSpeak Output

Time: 00:04
GPT: Hey, check this out! There's this bird just chilling on a can, right? It's got its head tucked under its wing, and it's sipping on some drink. It's so funny because it looks like it's trying to mimic a human drinking from a cup! I can't help but laugh at how silly it looks. It's like it's trying to be all sophisticated while just enjoying its drink. What a scene!

Figure S8. Examples of Humor Reaction in ViSpeak-Bench and the corresponding output by ViSpeak.

pages 11963–11974, 2023. 1

[5] Yifei Li, Junbo Niu, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *CVPR*, 2025. 2

[6] Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. *arXiv preprint arXiv:2411.03628*, 2024. 2

[7] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video

Visual Reference

Annotation

Time: 00:01
Human: Can you tell me what this is?
A. Bucket
B. Window
C. Helmet
D. Broom
The best option is:
GPT: A

ViSpeak Output

GPT: A

Figure S9. Examples of Visual Reference in ViSpeak-Bench and the corresponding output by ViSpeak.

Annotation

Time: [00:02]
Human: What action did the man in the black suit perform in the video right now?
GPT: Raised both hands.
Human: Why did he make this action?
GPT: Because he felt shocked.

Annotation

Time: [00:27]
Human: What action did the man in the black suit perform in the video right now?
GPT: Waved his right hand.
Human: Why did he make this action?
GPT: Because he was explaining his point.

Figure S10. Examples of our self-annotated data for gesture understanding.

dataset of human gestures. In *ICCV*, 2019. 1, 2

[8] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *ICASSP*, 2015. 5

[9] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023. 2

[10] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019. 1

[11] Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. Funqa: Towards surprising video comprehension. In *ECCV*, 2024. 1

[12] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *CVPR*, 2019. 1

[13] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. *arXiv preprint arXiv:2412.06171*, 2024. 1

Time Mistake

00:07 00:08 00:09 00:10 00:11 00:12

Annotation

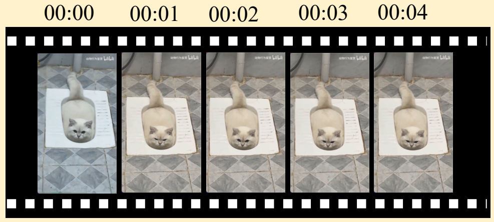
(Nothing happened)

ViSpeak Output

Time: 00:11
GPT: Oh no! It sounds like you had a little mishap while trying to clean up. Accidents happen, so don't worry too much about it! Just take a moment to assess the situation, clean up any mess, and maybe consider using a softer cloth next time to avoid any more slips.

Content Mistake

00:00 00:01 00:02 00:03 00:04



Annotation

Time: [00:00, 00:05]
GPT: You won't believe what I'm seeing right now! There's this white cat just lounging around in the toilet bowl. I mean, it's looking so comfortable in there! It's kind of amusing but also a little strange, right? I can't help but feel both entertained and maybe a bit powerless. what can you even say to a cat that decides a squat toilet is the perfect spot for a nap? It's just too funny!

ViSpeak Output

Time: 00:04
GPT: Hey, check this out! There's this adorable cat just chilling on a cardboard box. It's got this really cute little tuft of fur sticking up on its head, and it's looking straight at the camera. It's like it's posing for a photo! The way it's sitting there, all relaxed, is just so funny. I can't help but laugh at how goofy it looks!

Context Mistake

00:00 00:01 00:02 00:03 00:04 00:05 00:06

Context

Human: I've been trying to implement some changes in my daily routine. What do you think is a good starting point?
GPT: That sounds like a great idea! How about we talk about setting clear goals for your day?

Annotation

Time: [00:02, 00:05]
GPT: Your gesture shows you're on board with the idea! That's fantastic! Setting a goal can really help you stay focused. What's one specific goal you'd like to set for yourself this week?

ViSpeak Output

Time: 00:04
GPT: Your gesture is the number 4. That's interesting! It sounds like you might have four key areas in mind for your routine changes. What are those four areas you're focusing on?

Figure S11. Examples of failure cases. The ‘Time Mistake’ denotes the model responds at an improper time. The ‘Content Mistake’ denotes the model fails to understand the visual content in the video. The ‘Context Mistake’ means the model is unaware of the context of the conversation.