

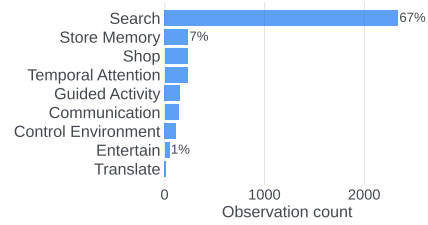
A Dataset Statistics

Figure 1 visualizes distributions over several dimensions of our dataset, including:

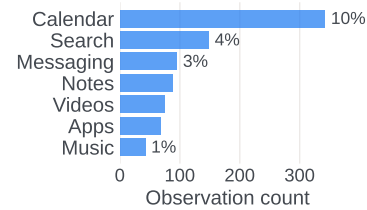
- (Fig. 1a) **Goal types**. We can see that our dataset is skewed towards “Search” goal types (~2 thirds of all observation-goal pairs), given their generality and suitability.
- (Fig. 1b) **Script Description Word Cloud**. We took the short descriptions of each script (e.g., “Troubleshoot a non-functioning leaf blower”) and plotted a word cloud showing the top 250 words, sized proportionally to the log of the count of the word. Themes emerge, such as memory, learning, health and fitness, meal preparation, daily chores, and recreation.
- (Fig. 1c) **Digital apps**. “Calculator” app is the most used in across observations (~10%), followed by “Search” (~4%) and “Messaging” (~3%).
- (Fig. 1d) **Density and diversity of recordings**.
 - (i) In terms of scripts, we can see from the histograms that the majority of scripts were recorded by 5-7 participants, while a small minority were recorded by as many as 15 participants. On the other hand, the distribution of number of videos per script is more uniform ranging from 1 to ~40 videos per script.
 - (ii) In terms of participants, we observe a concentration around 2-3 scripts per participant, while some participants recorded as many as 10 scripts. We also observe a concentration in participants who recorded <10 videos, while a small minority recorded as much as 40 videos.
- (Fig. 1e) **Modality volumes**. We compute histograms of observation volumes for all modalities. We observe histograms that exhibit bell-like shaped curves when plotted in logarithmic scale.
 - V: video durations in seconds for vision,
 - A: word counts for audio (we discard non-speech audio),
 - D: digital states in kilobytes (KB)
 - L: history data (vision + audio + digital) in kilobytes (KB)

We note that the video dataset we initially collected is 264 hours in length. When considering only videos that passed an initial quality review (considering the video, audio, and generated digital state quality), the size is reduced to 155 hours. Finally, applying the context-windowing (which eliminates context that is off-script) further reduces it to 29 hours.

App Type	Description	Arguments
Search	Web or timeline search	source, query_item, query
Shop	Shopping	content
Store Memory	Grocery lists, todo list etc	content
Temporal Attention	Timers and reminders	action, time, content
Guided Activity	Workout videos, tutorials, etc	content, time
Communication	Messaging	action, target, content
Control Environment	IoT (lights, smart speaker, etc)	target, value
Translate	Audio or text translation	modality, src_language
Entertain	Playing music or videos	modality, content

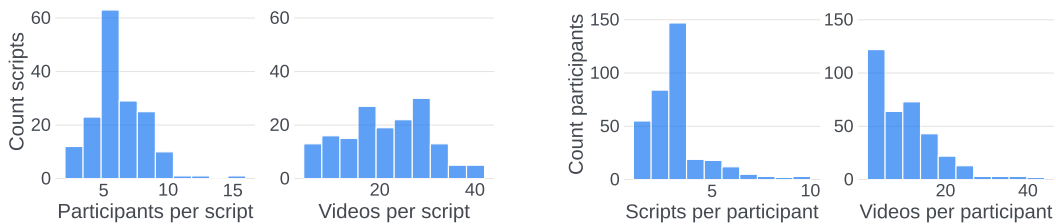


(a) (left) Table describing the types of digital goals in the dataset, and listing the arguments associated with each. (right) The number of videos of each type of goal in the dataset.

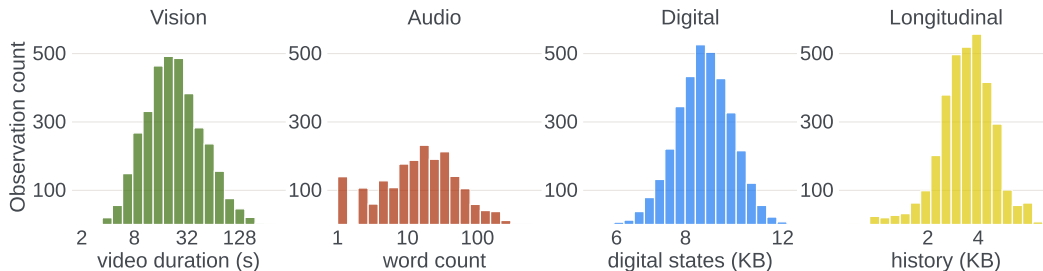


(b) Word cloud of script descriptions.

(c) Distribution of app annotations for generating digital context.



(d) Distributions of participants and videos per script (left), and of scripts and videos per participant (right).



(e) Distributions over modality volumes (per-observation), in logarithmic scale (base-10 for audio and base-2 for the rest).

Figure 1: Overall dataset statistics

All examples are rated as very relevant

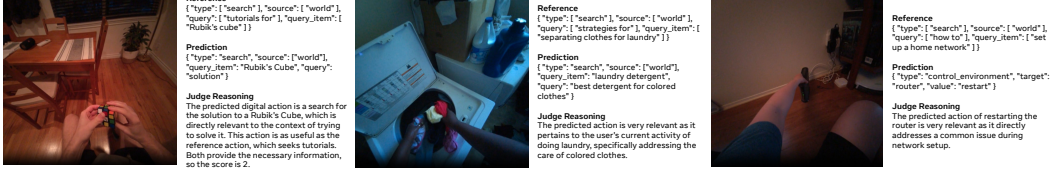


Figure 2: Goal inference examples with vision-only contexts. Best viewed when zoomed in.

B Qualitative Results

B.1 Generative Examples

In this section we show examples of goal inference from various modalities and analyze where different input modalities help or fail.

B.1.1 Vision-only Context Examples

In Fig. 2, we present several examples showing the model’s (Qwen2.5-VL-72B) prediction using only vision-contexts (i.e. video frames). All three examples are drawn from S_V subset which means supposedly these goals can be inferred with vision contexts only. With no surprise, all three predicted goals are considered very relevant by the LLM Judge, despite for the center and right example, they differ from the reference goals.

B.1.2 Audio Context Examples

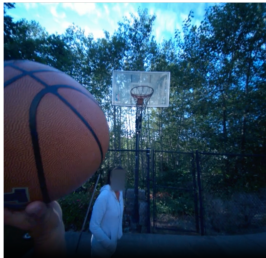
Fig. 3 shows three examples from the subset S_{VA} where Qwen2.5-VL-72B’s predictions with V inputs are irrelevant, while predictions with VA inputs are very relevant. These examples, which feature varied locations and lighting settings, illustrate the importance of the audio modality in cases where vision alone can be very misleading due to the lack of relevant contextualized information present in audio transcriptions. The audio transcriptions shown in the figure may be subject to spelling errors due to limitations in Automatic Speech Recognition (ASR) system used for transcription.

B.1.3 Digital Context Examples

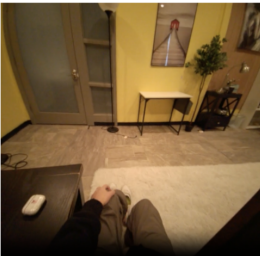
Fig. 4 shows three examples from the subset S_{VD} from Qwen2.5-VL-72B using multiple input modalities V, VD and VD*. In Fig. 4a, we see the typical case where visual and digital modalities are well-aligned and additional information in VD and VD* modalities helps predict the goal accurately over the V modality. Fig. 4b shows a failure case where both vision and digital states have many distractors and the model is not able to accurately predict the goal, except by using the high-signal VD* modality. Finally, Fig. 4c shows a case where the visual modality dominates the model’s prediction and even with VD*, it is hard to predict the correct digital goal as the model focuses on the incorrect visual distractor (Christmas tree in the background).

B.1.4 Longitudinal Context Examples

Fig. 5 shows three examples from the subset S_{VL} from Qwen2.5-VL-72B using multiple input modalities V, VL and VL*. In Fig. 5a, we see how longitudinal history assists VL and VL* modalities to personalize their user goal prediction relative to the V modality. In Fig. 5b, potentially due to the high amount of distracting user history in the VL modality, we see how the model is unable to accurately predict a personalized goal. Only the high-signal VL* modality that is void of distracting history observations uses both video-specific visual cues (about to run out of bread) and longitudinal history (of the user adding pasta to the grocery list) to recommend adding bread to the grocery list. Fig. 5c shows a case where the visual cues override longitudinal history and shows the model’s prediction being relevant to organizing a cluttered space. But, both VL and VL* input modalities fail to identify that the user likes to listen to music while de-cluttering their space.

<p>Reference Goal</p> <pre>{ "type": "search", "source": "world", "query": "find", "query_item": "chiropractor" }</pre> <p><i>Comment</i> Despite the prediction with V being very plausible, the inclusion of audio in VA results in the prediction of a higher priority goal.</p>	<p>V Modality</p> 	<p>A Modality</p> <p>00:03 - 00:05 [Person 1] Okay, here we are. That's what court.</p> <p>00:07 - 00:11 [Person 2] Ah! Are you okay? I think I really hurt myself badly.</p> <p>00:12 - 00:15 [Person 1] I should look for a chiropractor online and make an appointment.</p>
<p>Prediction with V</p> <pre>{ "type": "communication", "action": "share", "target": "social media", "content": "basketball moment with friend" }</pre> <p>Judge Score: 0.0</p>	<p>Prediction with VA</p> <pre>{ "type": "search", "source": "world", "query_item": "chiropractor near me", "query": "appointment availability" }</pre> <p>Judge Score: 1.0</p>	

(a) Outdoors example.

<p>Reference Goal</p> <pre>{ "type": "search", "source": "world", "query": "find", "query_item": "hiking trails" }</pre> <p><i>Comment</i> Prediction with V could work given its generality, but the prediction with VA is much more relevant given the context.</p>	<p>V Modality</p> 	<p>A Modality</p> <p>00:05 - 00:09 [Person 1] what it looks really good today. Let's go for a hike.</p> <p>00:13 - 00:17 [Person 1] I'm in the right air. I want to wear a shickia this time.</p> <p>00:18 - 00:21 [Person 1] The last hike was pretty good. The little bit too long for me.</p>
<p>Prediction with V</p> <pre>{ "type": "search", "source": "world", "query_item": "furniture arrangement ideas", "query": "inspiration" }</pre> <p>Judge Score: 0.0</p>	<p>Prediction with VA</p> <pre>{ "type": "search", "source": "world", "query_item": "hiking trails near me", "query": "suitable hiking trails for beginners" }</pre> <p>Judge Score: 1.0</p>	

(b) Indoors example with good lighting conditions.


<p>Reference Goal</p> <pre>{ "type": "search", "source": "world", "query": "how to", "query_item": "file taxes" }</pre> <p><i>Comment</i> The prediction with VA is much more relevant than just V, given the essential information in audio transcriptions.</p>	<p>V Modality</p> 	<p>A Modality</p> <p>00:02 - 00:05 [Person 1] Let's see. We have some tasks to do today. Task one.</p> <p>00:06 - 00:09 [Person 1] Check. Test two. Check. File taxes.</p> <p>00:10 - 00:13 [Person 1] How will I do that? I'm not sure how to file my taxes.</p>
<p>Prediction with V</p> <pre>{ "type": "entertain", "modality": ["audio", "vision"], "content": "relaxing music playlist" }</pre> <p>Judge Score: 0.0</p>	<p>Prediction with VA</p> <pre>{ "type": "search", "source": ["world"], "query_item": "filing taxes", "query": "how to file taxes" }</pre> <p>Judge Score: 1.0</p>	

(c) Indoors example with darker lighting conditions.

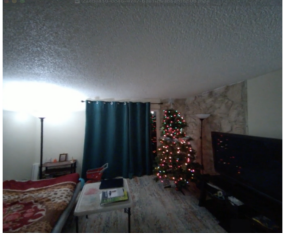
Figure 3: Qualitative examples from the audio modality.

<p>Reference Goal</p> <pre>{'type': 'search', 'source': 'world', 'query': 'nearby movie times for', 'query_item': 'inside out 2'}</pre> <p>Cues {'Calendar': 'several upcoming calendar events for going to see movies.', 'Videos': 'The user opens YouTube and plays inside out 2 trailer.'}</p> <p>Comment <i>V fails to identify the movie, but both VD and VD* are able to predict meaningful goals for the correct movie based on the digital state</i></p>	<p>V Modality</p> 	<p>D* Modality</p> <pre>'current_datetime': '2025-04-21 17:57'</pre> <pre>'Calendar': [..., {'title': 'Cinema Evening with Caleb Hall', 'start_datetime': '2025-04-23 18:00', 'end_datetime': '2025-04-23 21:00', 'location': 'Landmark Theatres, 911 Pine St, Seattle', ...}]</pre> <pre>'Videos': [..., {'title': 'Inside Out 2 Official Trailer', 'duration_secs': 147, 'category': 'Movies', 'watch_timestamp': '2025-04-20 20:45:00'}]</pre>
<p>Prediction with V</p> <pre>{'type': 'search', 'source': ['world'], 'query_item': 'Everything Movie', 'query': 'release date and cast'}</pre> <p>Judge Score: 0.0</p>	<p>Prediction with VD</p> <pre>{'type': 'search', 'source': ['world'], 'query_item': 'Inside Out 2', 'query': 'showtimes and ticket availability'}</pre> <p>Judge Score: 1.0</p>	<p>Prediction with VD*</p> <pre>{'type': 'search', 'source': ['world'], 'query_item': 'Inside Out 2', 'query': 'release date and cast'}</pre> <p>Judge Score: 1.0</p>

(a) Both VD and VD* correctly predict the user's goal

<p>Reference Goal</p> <pre>{'type': 'search', 'source': 'world', 'query': 'how to play', 'query_item': 'Clue Rivals Edition'}</pre> <p>Cues {'Calendar': 'The user sees a calendar event "playing Clue Rivals Edition with friend".}</p> <p>Comment <i>Too many distractor objects in vision (for V) and large digital state for VD. Only VD* has the exact digital context to predict meaningful goals.</i></p>	<p>V Modality</p> 	<p>D* Modality</p> <pre>'current_datetime': '2023-10-31 09:09'</pre> <pre>'Calendar': [{ 'title': 'Playing Clue Rivals Edition with Kaito', 'start_datetime': '2023-11-01 19:00', 'end_datetime': '2023-11-01 21:30', 'location': 'Game Center, Shinjuku-ku, Tokyo', 'attendees': ['Kazuki Nakamura', 'Kaito Matsumoto'] }]</pre>
<p>Prediction with V</p> <pre>{'type': 'entertain', 'modality': ['audio', 'vision'], 'content': 'watching movies or shows'}</pre> <p>Judge Score: 0.0</p>	<p>Prediction with VD</p> <pre>{'type': 'entertain', 'modality': ['audio', 'vision'], 'content': 'latest movie trailers'}</pre> <p>Judge Score: 0.0</p>	<p>Prediction with VD*</p> <pre>{'type': 'entertain', 'modality': ['audio', 'vision'], 'content': 'board game strategy videos'}</pre> <p>Judge Score: 1.0</p>

(b) Only VD* correctly predicts the user's goal

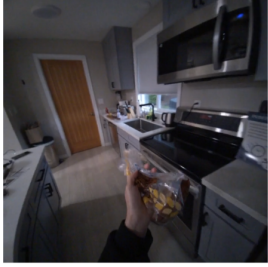
<p>Reference Goal</p> <pre>{'type': 'search', 'source': 'world', 'query': 'find', 'query_item': 'application info for University of Washington'}</pre> <p>Cues {'Calendar': 'The user has application deadline for UW in calendar.'}</p> <p>Comment <i>Fails because vision focuses on Christmas tree and a majority of information is hidden in audio</i></p>	<p>V Modality</p> 	<p>D* Modality</p> <pre>'current_datetime': '2024-06-16 11:23'</pre> <pre>'Calendar': [{ 'event_id': '2d7946ae967e4858a35ba73022452cf4', 'title': 'Finalize and Submit UW Application', 'start_datetime': '2024-06-19 09:00', 'end_datetime': '2024-06-19 11:23', 'location': 'Home Office', 'attendees': [] }]</pre>
<p>Prediction with V</p> <pre>{'type': 'search', 'source': ['world'], 'query_item': 'Christmas tree decoration ideas', 'query': 'inspiration'}</pre> <p>Judge Score: 0.0</p>	<p>Prediction with VD</p> <pre>{'type': 'search', 'source': ['world'], 'query_item': 'Christmas music', 'query': 'playlist'}</pre> <p>Judge Score: 0.0</p>	<p>Prediction with VD*</p> <pre>{'type': 'search', 'source': ['world'], 'query_item': 'Christmas tree decoration ideas', 'query': 'inspiration'}</pre> <p>Judge Score: 0.0</p>

(c) Neither VD nor VD* correctly predicts the user's goal

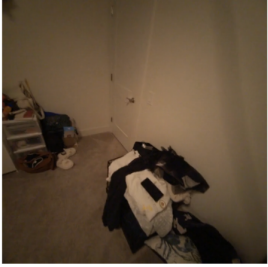
Figure 4: Qualitative examples from the digital modality

<p>Reference Goal</p> <pre>{'type': 'guided_activity', 'content': 'white noise', 'time': '0:01:48'}</pre> <p><i>Comment</i> VL and VL* combine user history and visual observations to accurately predict the goal.</p>	<p>V Modality</p> 	<p>L* Modality</p> <p>Context: The video captures a serene bedtime routine. The person adjusts the bedding and prepares for sleep. The lamp casts a warm glow, highlighting the room's cozy ambiance. The person wears a watch, indicating their presence in the scene.</p> <p>Action: {'type': 'guided_activity', 'content': 'white noise', 'time': '0:02:06'}</p>
<p>Prediction with V</p> <pre>{ "type": "search", "source": ["world"], "query_item": "lamp design", "query": "inspiration for modern lamp designs" }</pre> <p>Judge Score: 0.0</p>	<p>Prediction with VL</p> <pre>{ "type": "guided_activity", "content": "relaxation exercise", "time": "15 minutes" }</pre> <p>Judge Score: 1.0</p>	<p>Prediction with VL*</p> <pre>{ "type": "entertain", "modality": ["audio"], "content": "relaxing music" }</pre> <p>Judge Score: 1.0</p>

(a) Both VL and VL* correctly predict the user's goal

<p>Reference Goal</p> <pre>{'type': 'store_memory', 'content': 'add bread to grocery list'}</pre> <p><i>Comment</i> V and VL make generic predictions about retrieving food from the refrigerator. VL* enhances this by incorporating past history and current low bread stock to add items to the shopping list.</p>	<p>V Modality</p> 	<p>L* Modality</p> <p>Context: The setting includes typical kitchen items, highlighting a cooking environment. The video shows a kitchen scene with a stove, pots, and a cutting board. A hand interacts with organic food items on the counter.</p> <p>Action: {'type': 'store_memory', 'content': 'add pasta to grocery list'}</p>
<p>Prediction with V</p> <pre>{ "type": "search", "source": ["world"], "query_item": "banana", "query": "recipes" }</pre> <p>Judge Score: 0.0</p>	<p>Prediction with VL</p> <pre>{ "type": "store_memory", "content": "I retrieved a packaged item from the refrigerator and placed it on the counter" }</pre> <p>Judge Score: 0.0</p>	<p>Prediction with VL*</p> <pre>{ "type": "store_memory", "content": "add bread and milk to grocery list" }</pre> <p>Judge Score: 0.5</p>

(b) Only VL* correctly predicts the user's goal

<p>Reference Goal</p> <pre>{'type': 'entertain', 'modality': 'audio', 'content': 'song playlist'}</pre> <p><i>Comment</i> V makes irrelevant prediction. VL and VL* identify a cluttered space making relevant predictions. But they fail to personalize from history that the user likes to listen to music while decluttering.</p>	<p>V Modality</p> 	<p>L* Modality</p> <p>Context: In a dimly lit garage, a person moves around and inspects the area filled with parked cars and scattered items. The setting suggests a private residential space used for both parking and storage, highlighting a mundane, everyday scenario.</p> <p>Action: {'type': 'entertain', 'modality': 'audio', 'content': 'song playlist'}</p>
<p>Prediction with V</p> <pre>{ "type": "search", "source": ["world"], "query_item": "Automatic Storage Water Heater", "query": "installation instructions" }</pre> <p>Judge Score: 0.0</p>	<p>Prediction with VL</p> <pre>{ "type": "search", "source": "world", "query_item": "storage solutions", "query": "best ways to organize a cluttered home" }</pre> <p>Judge Score: 0.0</p>	<p>Prediction with VL*</p> <pre>{ "type": "search", "source": ["world"], "query_item": "home organization tips", "query": "ways to declutter and organize a home" }</pre> <p>Judge Score: 0.0</p>

(c) Neither VL nor VL* correctly predicts the user's goal

Figure 5: Qualitative examples from the longitudinal modality

B.2 Comparing Human to Model Performance via MCQ Task Examples

To qualitatively compare human and model performance, we present a few sets of MCQ examples from the human predictability study in Figures 6, 7, 8, 9. Each row represents one MCQ problem. A relevant frame from the video is shown on the left. On the right is a block of text containing (1) a *description* of what happens in the video, (2) any transcribed *speech* from the audio, (3) the average human and model (across all tested models) accuracies, (4) the *reference goal*, (5) (in green) MCQ options that humans selected, (6) (in purple) MCQ options that models selected, and (7) (in gray) the full set of MCQ options.

Figure 6 contains “easy” examples which both humans and models are able to predict with high accuracy. Figure 7 contains examples where humans appear to have strong intuitions about which goals may be relevant, but models fail. Figure 8 contains examples where there are *multiple relevant goals* in the option-set (examples where the strong reference bias of MCQ introduces noise into the evaluation). Figure 9 contains examples that require *fine-grained* visual recognition (e.g., reading text or identifying small objects like house plants) in order to solve, where models often struggle.

MCQ Examples Where Models Succeed

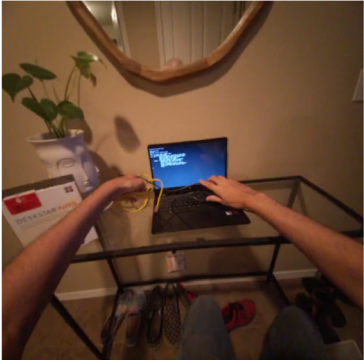
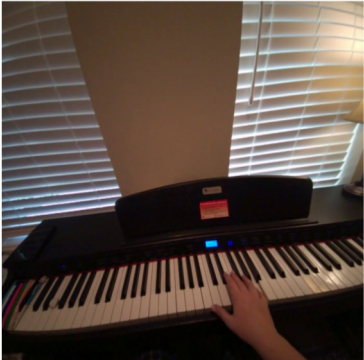

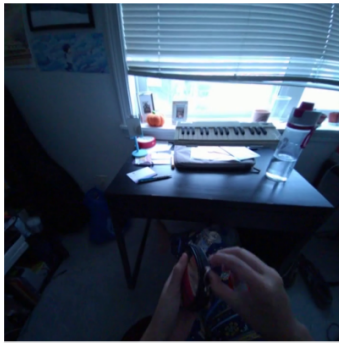
	<p><i>Description: (home network setup)</i></p> <p>Speech: "Okay..network.setup time. Over the last year or so, I've been doing more vide..."</p> <p>Human Acc: 1.000 Model Acc: 1.000</p> <p>Reference: Do a search for 'how to set up a home network'</p> <p>Human Predictions: ["Do a search for 'how to set up a home network'"]</p> <p>Model Predictions: ["Do a search for 'how to set up a home network'"]</p> <p>Option Set: ['Shop for body wash' "Do a search for 'find book in series after gone by midnight'" "Do a search for 'how to set up a home network'" "Translate audio from hindi to English"]</p>
	<p><i>Description: Tries to play a C chord on piano</i></p> <p>Speech: "This chord doesn't sound right..Hm..Shit algorithm piano..Hi..welcome to Roc..."</p> <p>Human Acc: 1.000 Model Acc: 1.000</p> <p>Reference: Do a search for 'how to play C on piano'</p> <p>Human Predictions: ["Do a search for 'how to play C on piano'"]</p> <p>Model Predictions: ["Do a search for 'how to play C on piano'"]</p> <p>Option Set: ["Do a search for 'find becoming'" "Control control my bedroom, set to off" "Remember I ate 8 total grams of protein" "Do a search for 'how to play C on piano'"]</p>
	<p><i>Description: Draws a vase on paper</i></p> <p>Speech: "I did it out to my design wall and then I started measuring people. A lot of ..."</p> <p>Human Acc: 1.000 Model Acc: 1.000</p> <p>Reference: Do a search for 'images of vases and flowers'</p> <p>Human Predictions: ["Do a search for 'images of vases and flowers'"]</p> <p>Model Predictions: ["Do a search for 'images of vases and flowers'"]</p> <p>Option Set: ["Do a search for 'information about best location for barrel cactus'" "Send a message to Stephen saying 'photo of flowers'" "Do a search for 'images of vases and flowers'" "Do a search for 'how to organize bookshelf'"]</p>

Figure 6: See Section B.2 in the text for a full description of the figure.

MCQ Examples where Models Struggle to Intuit Goals



Description: Person looks at laundry and checks coin purse

Speech: "there."

Human Acc: 1.000 Model Acc: 0.500

Reference: Do a search for 'location of laundromat'

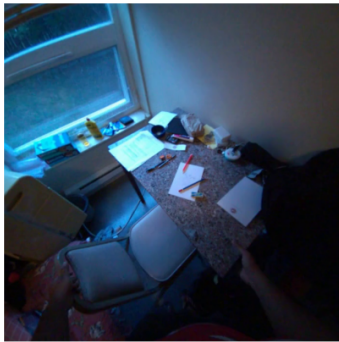
Human Predictions:

["Do a search for 'location of laundromat'"]

Model Predictions:

["Do a search for 'how to clean bagless vacuum cleaner'"
"Do a search for 'location of laundromat'"]

Option Set: ["Do a search for 'how to clean bagless vacuum cleaner'"
'Remember remember my car is located here'
"Do a search for 'information about indigo urgent care'"
"Do a search for 'location of laundromat'"]



Description: Person looks around their cramped space and tries to move a chair

Speech: "man..Stu long.has set to.congress.the good"

Human Acc: 1.000 Model Acc: 0.500

Reference: Do a search for 'how to setup small living space'

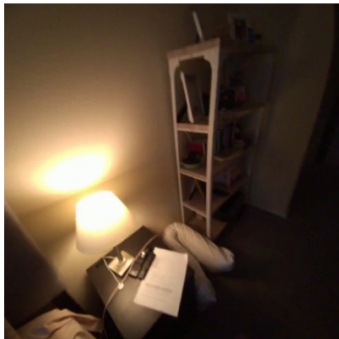
Human Predictions:

["Do a search for 'how to setup small living space'"]

Model Predictions:

["Do a search for 'how to setup small living space'"
"Do a search for 'how to install frame'" 'Control dining room, set to on']

Option Set: ["Do a search for 'how to setup small living space'"
"Do a search for 'how to install frame'" 'Control dining room, set to on'
"Do a search for 'strategies for separating clothes for laundry'"]



Description: Person looks at the lamp

Speech: ""

Human Acc: 1.000 Model Acc: 0.750

Reference: Control lamp, set to off

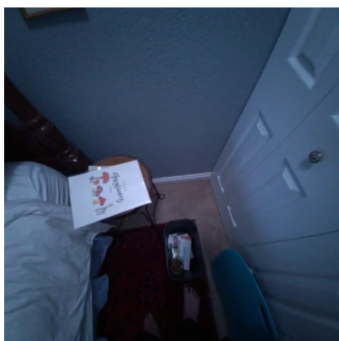
Human Predictions:

['Control lamp, set to off']

Model Predictions:

['Control lamp, set to off'
"Send a message to Brianna Clark saying 'Ed Sheeran - Shape of You'"]

Option Set: ["Do a search for 'recipes with marinara sauce, spaghetti, butter'"
"Send a message to Brianna Clark saying 'Ed Sheeran - Shape of You'"
'Launch a guided activity to do run near current location with a duration of 0:16:40'
'Control lamp, set to off']



Description: Person looks at recycling container

Speech: ""

Human Acc: 0.667 Model Acc: 0.500

Reference: Do a search for 'information about recycling schedule'

Human Predictions:

["Do a search for 'information about recycling schedule'"
'Control control my bedroom, set to off']

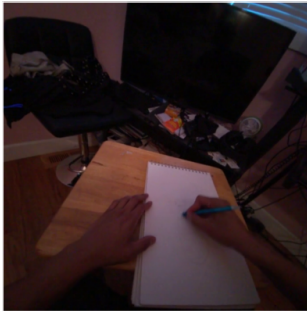
Model Predictions:

['Remember 3 cheese nachos'
"Do a search for 'information about recycling schedule'"
"Do a search for 'information about Justin Jefferson'"]

Option Set: ["Do a search for 'information about Justin Jefferson'"
'Control control my bedroom, set to off' 'Remember 3 cheese nachos'
"Do a search for 'information about recycling schedule'"]

Figure 7: See Section B.2 in the text for a full description of the figure.

MCQ Examples with Multiple Relevant Goals



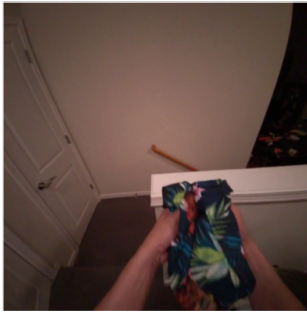
Description: Drawing a vase, but hard to make out the details. (Speech is hallucinated)
 Speech: "Being able to keep increasing the speed and increasing the speed is fine, the..."
 Human Acc: 0.000 Model Acc: 0.750

Reference: Do a search for 'images of vases and flowers'

Human Predictions:
 ["Do a search for 'tutorials for Drawing'"]

Model Predictions:
 ["Do a search for 'tutorials for Drawing'"
 "Do a search for 'images of vases and flowers'"]

Option Set: ["Do a search for 'tutorials for Drawing'"
 "Send a message to Stephen saying 'photo of flowers'"
 "Do a search for 'information about best location for barrel cactus'"
 "Do a search for 'images of vases and flowers'"]



Description: Person picks up a shirt and examines it

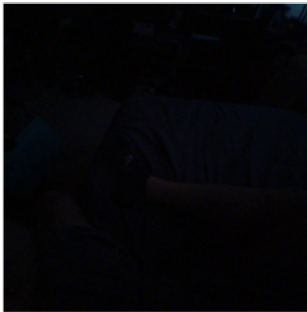
Speech: ""
 Human Acc: 0.667 Model Acc: 0.000

Reference: Do a search for 'how to hand wash delicate clothing'

Human Predictions:
 ["Remember Add fold laundry to todo list (complete by 8:00 PM)"
 "Do a search for 'how to hand wash delicate clothing'"]

Model Predictions:
 ["Remember Add fold laundry to todo list (complete by 8:00 PM)"]

Option Set: ["Do a search for 'how to hand wash delicate clothing'"
 "Do a search for 'how to display utensils'" "Shop for body wash"
 'Remember Add fold laundry to todo list (complete by 8:00 PM)']



Description: Person wakes up and complains of being tired

Speech: "Oh boy, I still feel pretty tired."
 Human Acc: 0.667 Model Acc: 0.000

Reference: Launch a guided activity to do Analyze sleep patterns with a duration of 10 minutes

Human Predictions:
 ["Launch a guided activity to do Analyze sleep patterns with a duration of 10 minutes"
 "Set a timer/reminder for 30 with the note 'Wake up timer'"]

Model Predictions:
 ["Set a timer/reminder for 30 with the note 'Wake up timer'"]

Option Set: ['Launch a guided activity to do guided meditation with a duration of 0:00:00'
 "Set a timer/reminder for 30 with the note 'Wake up timer'"
 'Launch a guided activity to do Analyze sleep patterns with a duration of 10 minutes'
 'Launch a guided activity to do run near current location with a duration of 0:16:40']



Description: Person rubs the carpet indicating a stain

Speech: ""
 Human Acc: 1.000 Model Acc: 0.750

Reference: Do a search for 'how to remove stain'

Human Predictions:
 ["Do a search for 'how to remove stain'"]

Model Predictions:
 ["Do a search for 'how to troubleshoot broken vacuum cleaner'"
 "Do a search for 'how to remove stain'"]

Option Set: ["Do a search for 'how to troubleshoot broken vacuum cleaner'"
 "Do a search for 'how to cut melon'"
 "Do a search for 'how to display plushes'"
 "Do a search for 'how to remove stain'"]

Figure 8: See Section B.2 in the text for a full description of the figure.

MCQ Examples where Models Struggle with Perception

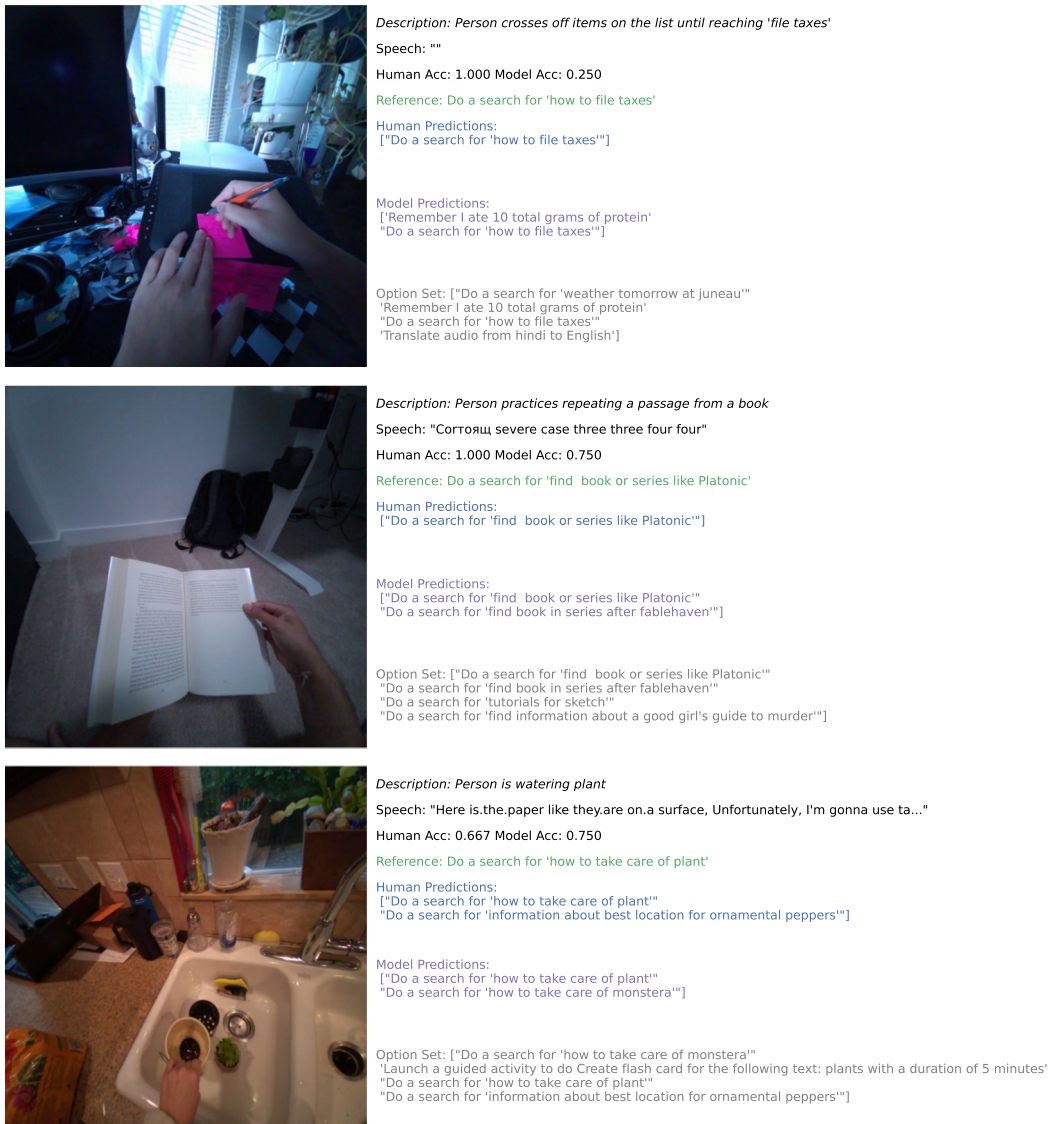
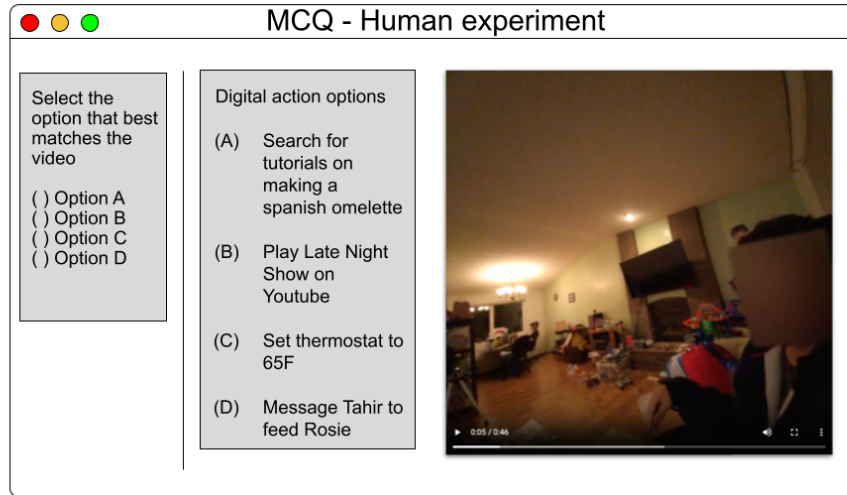


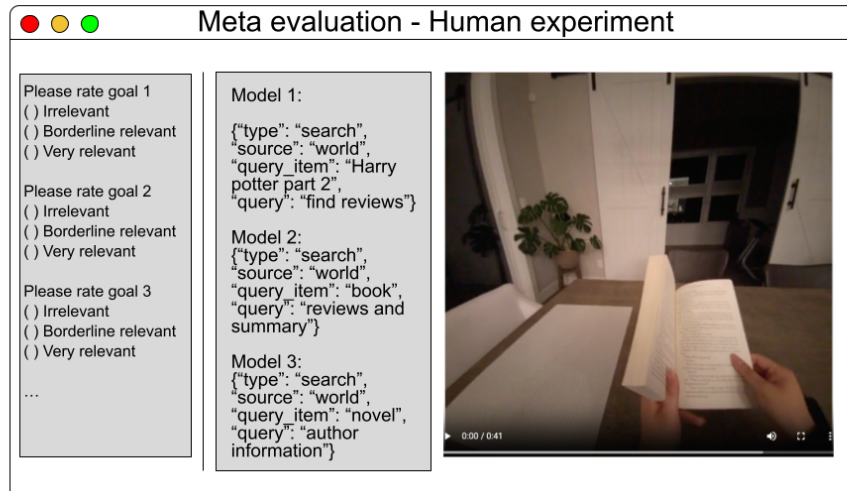
Figure 9: See Section B.2 in the text for a full description of the figure.

C Human experiment trial structure

The figures below show the trial design corresponding to a single trial of the MCQ human study (Fig. 10a) and the meta-evaluation human study (Fig. 10b).



(a) Example structure of a single trial in the human MCQ experiment



(b) Example structure of a single trial in the human meta evaluation experiment

D Modality-Specific Details

D.1 Digital Context Generation

As mentioned briefly in main text, we designed a pipeline for generating rich digital contexts representing the internal app states of seven widely-used apps: *Calendar*, *Messaging*, *Notes*, *Search*, *Videos*, *Maps*, and *Music*.

For this, we associated the relevant contextual cues of *digital* modality for each scenario with an app from the above seven apps. This resulted in 825 observation-goal pairs across 43 scenario scripts having at least one digital cue annotated with one of the seven apps. We show the resulting annotated app distribution in Figure 1c. For the remaining 2,652 observation-goal pairs without relevant digital cues, we generate digital context without any cue-conditioning.

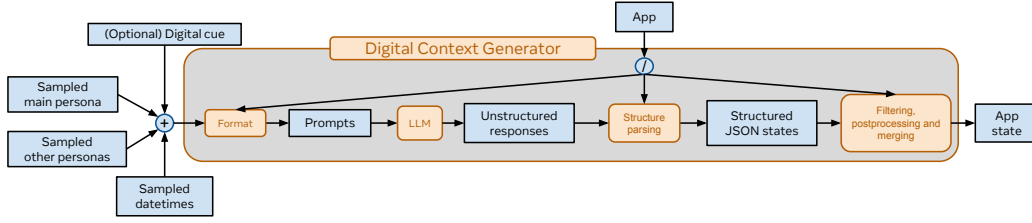


Figure 11: Digital Context Generator pipeline: A sampled main user persona, other personas, current datetime, and optionally a digital cue are processed via the DCG to generate structured app states.

The core of this system is a module we term the **Digital Context Generator (DCG)**, which synthesizes app-specific internal states from high-level persona and digital cues using a Large Language Model (LLM). Towards this we build a collection of 50 human personas containing synthesized names, ages, genders, nationalities, occupations and other useful fields. As illustrated in Figure 11, the DCG takes as input a sampled main user persona, 5 other related personas (e.g., friends, co-workers), a sampled current datetime and optionally a contextual digital cue. The process to generate the app states follows these stages:

1. **Formatting:** Inputs are filled into prompt templates with instructions to generate digital states for a given app. The prompt templates are crafted to reflect realistic app usage patterns. The input personas and the sampled datetime help to generate app states uniquely tailored to the scenario under consideration. If an optional digital context cue is provided, it is treated by a separate prompt for the relevant app since encoding a cue can require following specialized instructions for each app.
2. **LLM Inference:** The prompts are fed to an LLM (we used *Llama3.3-70b-instruct* model for this purpose), which outputs unstructured text describing plausible digital activity and app interactions for the user.
3. **Structure Parsing:** The unstructured LLM output is then parsed into structured JSON representations specific to each app’s internal data structures.
4. **Filtering, Postprocessing and Merging:** Structured states are filtered to reject or correct invalid values for fields of the internal data structures. If a digital cue was provided to an app, that app separately generates sub-states derived from the digital contextual cue. These are merged with the app sub-states generated without the cue into a unified data structure. This final output is a coherent snapshot of the user’s internal app state at the sampled time.

This approach enables the generation of diverse and human-like app states across a variety of temporal, personal, and situational contexts. These app states are semantically coherent, temporally relevant, and contextually grounded in the persona’s attributes and environment. We produce a realistic digital state for all apps for all the 3,477 videos in the benchmark even if they do not have an associated digital cue. This results in an average of about 8.9k characters of digital state input for each video for a total of about 31 million digital state characters across the full dataset. Note that this implies that digital context contains a lot of distractor information, which models must ignore to correctly infer the user’s goals.

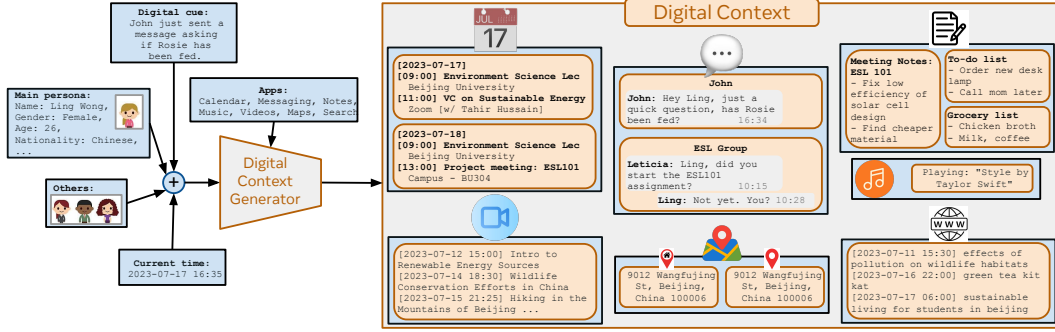


Figure 12: Example digital app states generated by the DCG: The main persona is “Ling”, a Beijing University student, with interests in environmental science and sustainable energy. The digital state reflects a snapshot of Ling’s app usage on a weekday evening. Note that the digital cue *John just sent a message asking if Rosie has been fed* is reflected in Ling’s Messaging app state.

An example of a fully generated digital state is shown in Figure 12 and demonstrates the DCG’s capacity to produce complex digital traces that mirror real-world app states.

D.2 Vision and Audio Processing Details

For all models except Llama models, we uniformly sample 32 frames from the context video as the input to the model. For Llama-3.2, we sample 1 frame since it’s only trained for single-image understanding. For video frame resolutions, we follow the default as much as possible which leads to input resolution 448x448 for GPT-4.1 and InternVL models, 700x700 for Qwen2.5 models and 1120x1120 for Llama-3.2, respectively.

To generate audio contexts, we use Whisper-base model [2] to transcribe all videos with a beam size of 5, and a temperature schedule of 0, 0.2, 0.4, 0.6, 0.8, and 1.0. We separately perform speaker diarisation and voice activity detection using internal models and add them to the transcription. Meanwhile, we pre-process the dataset to automatically generate captions for all videos using a set of VLMs, and perform filtering by checking the consensus of different VLM models. These video captions are then used as the longitudinal history contexts for evaluation.

D.3 Longitudinal History Details

In this section we elaborate on the procedure to generate history context cues for the longitudinal scenarios as defined in the main paper.

History Bank Generation. For each video corresponding to a longitudinal scenario, we sampled a history bank of 5 support videos from recordings of the same participant and environment. One of these support videos shares the same script as the test video; this corresponds to positive longitudinal history. The other videos which don’t share the test video’s script are distractors that are used to reflect an ecologically valid longitudinal history setup.

Textual representation of audiovisual history context We represent longitudinal history in the form of detailed video captions and audio transcriptions of support videos. video captions as longitudinal history contexts. For each 10-second clip of a support video, first we use Qwen2.5-VL-72B and InternVL2.5-78B-MPO to separately generate generate video summarization. We then leverage DeepSeek-R1-Distill-Llama-70B to process the two generations and merge them into one video summary. The LLM merger is instructed to remove any information that exists only in the generation from one models, and only keep the description that are shared across both VLM models. Once we obtain summaries for all 10s chunks of the support video, we concatenate those that fall within the context window annotation as a representation of the visual support context from this video. We note it is possible that this captioning process introduces noise into the representation of longitudinal history videos. However, we note that our use of the abovementioned LLM merger reduces the likelihood of such false information occurring in the captions. In addition, we also include the audio transcription of the video in the history.

Longitudinal evaluation of VLMs As shown below in Section F.1, each support video’s history context and structured goal annotation are packaged into a JSON dictionary and shuffled along with other support videos for a given test video during evaluation.

E Meta-Evaluation Implementation Details

In this section, we provide the implementation details for our Meta-Evaluation experiments designed to study the alignment between LLM Judge models and human raters.

For each generated goal, we assign three raters to assess its quality. Among them, results from the first two raters are used to obtain the “ground-truth” pairwise relative ranking order, while the third one is held out and later used to compute human-human alignment (we’ll elaborate this next). We loop over all possible pairs of predicted goals given a single video and filter out pairs of goals for which two human raters do not agree on their relative order (e.g., $<$ and $>$ are considered disagreeing each other, whereas $=$ and $<$ are not). We choose to filter goal pairs instead of individual goals as it’s much easier for humans to agree on the relative ordering of goals compared to giving consistent absolute scores. With this filtering, we end up removing 2.3% of model pairs where their rankings are inconsistent between raters. The filtered set is then used as ground-truth to compute ranking accuracy for judge models.

Specifically, for each video, we loop over all pairs in the filtered ground-truth set, and compute the ranking accuracy for the LLM Judge against the ground-truth relative ranking. For scenarios when LLM judge gives tied ranking to two VLM outputs, we consider the LLM judge is aligned with ground-truth if at least one human raters give the tied ranking. When using sentenceBERT as the judge, we treat two VLM outputs as tied if their sentenceBERT score difference is within 0.1. Finally, to measure the human-human alignment, we take the rating from the third human rater and compute its alignment to ground-truths in the same way as we did for LLM Judge models.

F Modeling Details

F.1 Example Prompts for Goal Inference and LLM Judge

First, we present the prompt for VLMs to answer multi-choice questions.

MCQ Prediction Prompt

```
You are an intelligent assistant. You will be given a video and four options, each
containing a digital action formatted as a dictionary. Your task is to respond with
which option (A or B or C or D) is the most likely digital action that follows a
human’s context window of actions present in the video.
```

```
{mcq_options}
```

NOTE:

- Your response should contain only the option letter A, B, C, or D. Only respond with one letter. Do not repeat the option.
- Wrap your final answer with `<answer>` and `</answer>` tag. For instance, an example full output should look like this: `<answer>A</answer>`

In addition to this base MCQ prompt, we optionally augment it with various modality contexts (i.e., audio, digital, longitudinal). Below we show examples for each of these context modalities.

Audio Contexts Example

```
To facilitate the task, the transcription for the video is provided as follows.
One of the speakers may be wearing smart glasses. Feel free to ignore the transcription
if it is not relevant. The transcription is:
```

```
Speaker 0: Let’s see. We have some tasks to do today. Task one.
Speaker 0: Check. Test two. Check. File taxes.
Speaker 0: How will I do that? I’m not sure how to file my taxes.
```

Digital Contexts Example

```
The current datetime is: ['2023-10-07 12:05']
```

```
Current state of the user’s Calendar app:
```

```
[...,
{'event_id': ['30a546df10154a99a302ca37d12d0b97'],
'title': ['Playing Cards Against Humanity with friend'],
'start_datetime': ['2023-10-07 20:00'],
'end_datetime': ['2023-10-07 23:00'],
'location': ["Javier’s home, 3456, Calle de la Luna, Madrid, Spain 28004"],
```

```

...
]

Current state of the user's Messaging app:
[{'participants': [('Javier Garcia'), ('Alexander Brooks')],
  'messages': [
    {'sender': ['Javier Garcia'],
      'message_id': ['6b228dbb2c54412f805ae28f7e76ac3b'],
      'timestamp': ['2023-09-20 10:00'],
      'content': ["Alexander, hope you're doing well. I've been meaning to ask, have you
        had a chance to read 'The Shadow of the Wind' by Carlos Ruiz Zafón? I think you'd
        appreciate the themes of literature and mystery."]},
    {'sender': ['Alexander Brooks'],
      'message_id': ['27fclcdbf4ac47b5a0431103cf779ea8'],
      'timestamp': ['2023-09-22 15:30'],
      'content': ['Javier, thanks for the recommendation! I actually just finished reading
        it and loved it. The way Zafón weaves the story around the Cemetery of Forgotten
        Books is captivating.']}
  ],
  'title': ['Alexander Brooks'],
  'conversation_id': ['f112cec95ee0448ea5301c71a7077124']}
],
...
]

Current state of the user's Search app:
[{'query': ['Spanish grammar exercises for elementary school'],
  'query_type': ['educational'],
  'timestamp': ['2023-09-20 14:30:00'],
  'query': ['Don Quixote summary and analysis'],
  'query_type': ['literary'],
  'timestamp': ['2023-09-20 14:45:00'],
  ...
]

Current state of the user's Videos app:
[{'title': ['Analysis of Don Quixote by Miguel de Cervantes'],
  'duration_secs': tensor([3600]),
  'tags': [('Spanish literature'), ('Don Quixote'), ('Miguel de Cervantes')],
  'category': ['Education'],
  'watch_timestamp': ['2023-09-20 20:30:00'],
  'title': ['How to Play Flamenco Guitar for Beginners'],
  'duration_secs': tensor([2700]),
  'tags': [('Music'), ('Flamenco'), ('Guitar')],
  'category': ['Music'],
  'watch_timestamp': ['2023-09-22 19:45:00'],
  ...
]

Current state of the user's Notes app:
[...
  {'title': ['Paella Recipe'],
    'content': ['Ingredients: saffron, chicken, seafood, rice. Instructions: heat oil, add
      onion and garlic, add rice and cook until lightly toasted. Add broth and simmer until
      rice is cooked.'],
    'last_updated': ['2023-09-20 18:00'],
    'title': ['Lesson Plan To-Do List'],
    'content': ['Prepare slides for Don Quixote, create reading comprehension questions,
      assign group project on character analysis. Due: 2023-10-12'],
    'last_updated': ['2023-10-05 10:00'],
    ...
  ],
  {'title': ['Grocery Shopping List - Paella Ingredients'],
    'content': ['Saffron, chicken breast, shrimp, mussels, rice, onion, garlic, olive oil,
      seafood broth'],
    'last_updated': ['2023-10-03 12:00']}
]

Current state of the user's Maps app:
{'home_location': ['3456, Calle de la Luna, Madrid, Spain 28004'],
  'current_location': ['3456, Calle de la Luna, Madrid, Spain 28004']}

Current state of the user's Music app:
{'currently_playing': ['']}

```

Longitudinal Contexts Example

Here is a description of the user's current context: ["The video captures a person in a e garage interacting with a bicycle, with the garage door open to reveal a driveway and parked cars outside."]

Here are a list of the user's past actions with corresponding text context to help you in this task:

```
{
  "context": [
    "In a bathroom, a person in a patterned robe uses a smartphone near a sink. The scene includes a mirror reflecting the room and decorative plants. In a bathroom, a person in a floral-patterned top interacts with a soap dispenser, holding it near the sink. The scene includes a decorative plant and a towel, emphasizing a personal care routine."
  ],
  "action": [
    "{ 'type': 'shop', 'content': 'hand soap' }"
  ],
  "transcription": "Speaker 0: Do you\nSpeaker 0: know what we\nSpeaker 0: want? Do you know what we want? Do you know what we want? Do you"
}

{
  "context": [
    "The video shows a person in a dimly lit room using a smartphone, possibly controlling or interacting with a stereo system. The room contains various electronic devices and furniture, with the person moving towards the stereo while holding the phone. The video shows a person interacting with a gaming setup in an indoor space. The setup includes a console or computer, controllers, and furniture such as a red couch or chairs. A fan is present, and the room is moderately lit."
  ],
  "action": [
    "{ 'type': 'temporal_attention', 'action': 'set', 'time': '1:00:00', 'content': 'stop gaming' }"
  ],
  "transcription": ""
}

{
  "context": [
    "In a nighttime setting, a person interacts with a smartphone near a parked blue car and a bicycle, preparing to ride. The video captures a calm bicycle ride through a quiet suburban area, highlighting the smooth movement and the serene environment with minimal distractions."
  ],
  "action": [
    "{ 'type': 'control_environment', 'target': 'bike recording app', 'value': 'start' }"
  ],
  "transcription": ""
}

{
  "context": [
    "In a bathroom, a person interacts with a tube near a sink and mirror. The scene includes a white sink, a mirror, and a gray shirt with 'Windows.NET' text. The person holds a tube with a red label. In a warmly lit bathroom, a person squeezes a white and red tube. The scene includes a sink, mirror, and various bathroom items, with the person wearing a gray T-shirt."
  ],
  "action": [
    "{ 'type': 'shop', 'content': 'toothpaste' }"
  ],
  "transcription": ""
}
```

Next, we provide the prompt used to generate digital goals.

Goal Generation Prompt

You are an intelligent agent.
Your task is to answer the following question:

Based on the images provided representing what the user is seeing, what digital action might they want to do on their phones?

To answer the question above, you MUST PICK A DIGITAL ACTION from the following TEMPLATE:

```
{
  "type": "search",
  "source": ["world", "timeline"],
  // "world" = general knowledge (facts, news, weather, etc)
  // "timeline" = user's history or environment (e.g., saved events)
  "query_item": str,           // e.g., "Red delicious apple"
  "query": str                 // e.g., "Nutritional content"
}

{
  "type": "store_memory",
  "content": str               // e.g., "I took my vitamins today"
}

{
  "type": "temporal_attention",
  "action": ["set", "unset"],
  "time": str,                // e.g., "Ten minutes"
  "content": str              // e.g., "Get ready for work"
}

{
  "type": "guided_activity",
  "content": str,              // e.g., "guided meditation"
  "time": str                 // e.g., "ten minutes"
}

{
  "type": "control_environment",
  "target": str,               // e.g., "smart lights in kitchen"
  "value": str                 // e.g., "brightness: 50%"
}

{
  "type": "communication",
  "action": ["message", "share"],
  "target": str,               // e.g., "Alex", "Instagram"
  "content": str
}

{
  "type": "translate",
  "modality": ["text", "audio"],
  "src_language": str          // e.g., "german"
}

{
  "type": "find_directions",
  "to": str                    // e.g., "nearest coffee shop"
}

{
  "type": "shop",
  "content": str               // e.g., "running shoes"
}

{
  "type": "entertain",
  "modality": ["audio", "vision"],
  "content": str               // e.g., "rock playlist", "cat videos"
}
```

Instructions:

- Choose only one action type that best fits the situation.
- Fill in all relevant fields in the template based on the visual context.
- DO NOT simply copy from the template. Write a meaningful, contextualized output.

We also present the prompt used by the LLM Judge to score generated goals.

LLM Judge Prompt

You are a judge. Your task is to evaluate the digital action predicted by the model. The digital action is a dictionary falling into one of the following goal types:

<SAME TEMPLATE AS IN GOAL GENERATION PROMPT>

Specifically, you are given a set of contexts that describe the video, as well as the predicted digital action that the user might want to take on their phones given the observed contexts.

You need to decide whether the predicted digital action is sensible given the contexts. To facilitate this, you are further given a reference digital action that is known to be sensible given the contexts.

Contexts:
{contexts}

Reference digital action:
{reference_digital_action}

Model predicted digital action:
{predicted_digital_action}

NOTE:

Please evaluate the prediction with a score that's either 0, 1 or 2 that corresponds to "irrelevant", "borderline relevant", or "very relevant".

A score of 0 → The digital action would probably not be useful for the person in the video.

A score of 1 → The digital action might be useful for the person in the video, but you're not confident.

A score of 2 → The digital action is definitely useful for the person in the video.

Please leverage the reference digital action when scoring the predicted digital action, but do not treat it as the single possible answer due to the multi-modal nature of this prediction task.

Please only output the score and nothing else. Do not add any explanation to your final answer.

Wrap the score with <score> and </score> tags. For instance, an example full output should look like this: <score>2</score>

Note that we used {0, 1, 2} judge scores in the prompt instead of {0, 0.5, 1.0} described in the paper, to use least amount of tokens (1 token rather than 3 comparing "1" vs "0.5"). We do post-processing to normalize them into {0, 0.5, 1.0}.

Finally, we provide the prompt used to generate video captions for longitudinal contexts. The prompt is adapted from [1].

Longitudinal Contexts Generation Prompt

MAIN INSTRUCTIONS:

Your task is to analyze video frames extracted from an Ego-centric video for a detailed video understanding exercise.

Examine the video frames closely and generate a comprehensive caption by strictly following the steps below:

Step 1: ****Scene Context****:

Observe the video. What is the primary setting and activity in the video?

Step 2: ****Spatial Relationship Analysis****:

Examine and report on the spatial relationships between key objects or characters in the video frames. Describe the positioning and orientation of each element relative to others.

Step 3: ****Detailed Object Analysis****:

List the key objects and characters in the frame. Describe their color, shape, texture, and any other notable features with precision. Focus on specific details like clothing, accessories, and colors.

Step 4: ****Motion Description****:

Identify and describe any significant motion or actions taking place.

Step 5: ****Text Analysis****:

Examine and return any significant texts observed in the video frames.

i.e.: menu, book title, billboards etc.

Step 6: ****Additional Details****:

Note any other important details or elements that stand out but are not covered by the above points, i.e.: gender, hair color, colors of accessories and other attributes in the video frames.

Step 7: ****Summary****:

Provide a concise yet comprehensive summary capturing the key elements and takeaways from this video following Steps 1 to 6 above. Your caption should encapsulate the scene's key aspects, offering a comprehensive understanding of its environment, activities and context.

GUIDELINES:

1. Strictly return your results in JSON format. Please see the example below:

```
```json
{
 "Scene Context": "A busy beach scene with families and surfers enjoying the sunny day.",
 "Spatial Relationship Analysis": "The sandcastle is in the foreground, the dog approaches from the left, and the surfer moves from right to center.",
 "Detailed Object Analysis": "Children are wearing colorful swimwear; the dog is a golden retriever; the surfer is wearing a blue and white wetsuit.",
 "Motion Description": "Children are building a sandcastle, a dog is running towards the water, and a surfer is catching a wave.",
 "Text Analysis": "There is a billboard on the beach with the ads that reads as Diving Equipment Rental",
 "Additional Details": "One child has red hair; the dog's leash is lying abandoned on the sand; multiple surfboards are visible in the background.",
 "Summary": "The video frames depict a joyful beach day emphasizing family activities, interaction with nature, and surfing as a key activity, showcasing the beach's vibrant atmosphere."
}
```

2. When not sure, please be conservative and restrain from adding uncertain information to your response. Do not make up facts. Do not make up information.

3. VERY IMPORTANT: YOU ARE ALLOWED TO USE A MAXIMUM OF 200 words in total.

## References

- [1] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 2024.
- [2] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL <https://arxiv.org/abs/2212.04356>.