Ego2Web: A Web Agent Benchmark Grounded in Egocentric Videos

Supplementary Material

7. More Experiment and Analysis

Fine-grained Domain Analysis. To further analyze domain-specific performance, Table 5 reports SR across five task categories. We observe that **BU-Gemini-3-Flash consistently outperforms all baselines across nearly every domain**, achieving the highest SR in E-Commerce (38.2%), Media Retrieval (50.7%), Knowledge Lookup (75.0%), and Local/Maps (48.3%). Across domains, **Knowledge Lookup** tasks are the easiest, with average SR reaching 50.0%, likely due to structured content and clearer objectives. In contrast, **Local/Maps** and **E-Commerce** are more challenging due to dynamic interfaces and multi-step interactions. Overall, these results suggest that (1) **multimodal grounding is essential** for solving visually rich web tasks and (2) **cross-domain generalization remains a key bottleneck**.

8. More Statistics of Ego2Web

Video Length. We select our egocentric videos based on EgoSchema [31] selection and trimming, which is constructed from the large-scale Ego4D [16] dataset. Ego4D contains over 3600 hours of first-person video with dense, timestamped narrations covering thousands of unique actions and objects. These narrations provide precise, fine-grained grounding of what the camera wearer is doing, making it a strong foundation for video reasoning tasks.

EgoSchema further refines this data to produce standardized clips suitable for reasoning and evaluation. One challenge in Ego4D is that videos vary widely in length and narration density. To ensure consistent temporal structure and sufficient semantic richness, EgoSchema filters for non-overlapping **three-minute clips** containing at least 30 human-annotated narrations. This creates segments with both stable duration and high-quality supervision, making them ideal for downstream multimodal reasoning, including our visually grounded web-task setting. We adopt these curated clips because their dense, structured narrations and standardized temporal format provide reliable visual cues and event structure that are essential for our benchmark.

Web Distribution. We provide a detailed breakdown of website distribution in Ego2Web in Table 6 and Figure 5. Our benchmark covers five major categories, including E-Commerce (230 tasks), Media Retrieval (132), Knowledge Lookup (92), Local/Maps (31), and Others (15), spanning 18 widely used websites. The distribution reflects realistic user behavior, where **e-commerce and media platforms dominate** real-world interactions. High-frequency domains such as Amazon, YouTube, and Wikipedia account for the majority of tasks, enabling rich multi-step reasoning and interaction scenarios. In contrast, lower-frequency categories (e.g., Local/Maps and Others) introduce long-tail but practically important environments such as Google Maps, Yelp, Reddit, and Booking, which involve diverse layouts and interaction patterns. This naturally imbalanced distribution poses additional challenges for web agents. While structured platforms (e.g., e-commerce and knowledge websites) provide relatively consistent interfaces, long-tail websites often exhibit higher variability in UI design and task structure, requiring stronger generalization ability. Overall, this diverse and realistic distribution ensures that Ego2Web evaluates agents across a wide spectrum of real-world web environments, preventing overfitting to a narrow set of websites and better reflecting practical deployment scenarios.

9. More Implementation Details

Instruction Generation Details. To enable realistic online evaluation, we collect a pool of popular and actively maintained websites. We then ask GPT-5 to generate web-task instructions conditioned on the detailed video captions and selected websites.

Agent Implementation Details. For agents who can not access video input, including Claude series and GPT-5.4, we convert video into structured, detailed video captions generated by strong MLLM (Gemini-3.1-Pro) to capture video details as much as possible and then feed the captions into the agents. We set the maximum step as 40 for each agent.

Human Evaluation. For human evaluation, three annotators assess each agent's output, and we apply majority voting to determine the final result.

Prompt Details. We provide detailed prompts for video captioning with QWen3-VL and the web task instruction generation, as shown in the following.

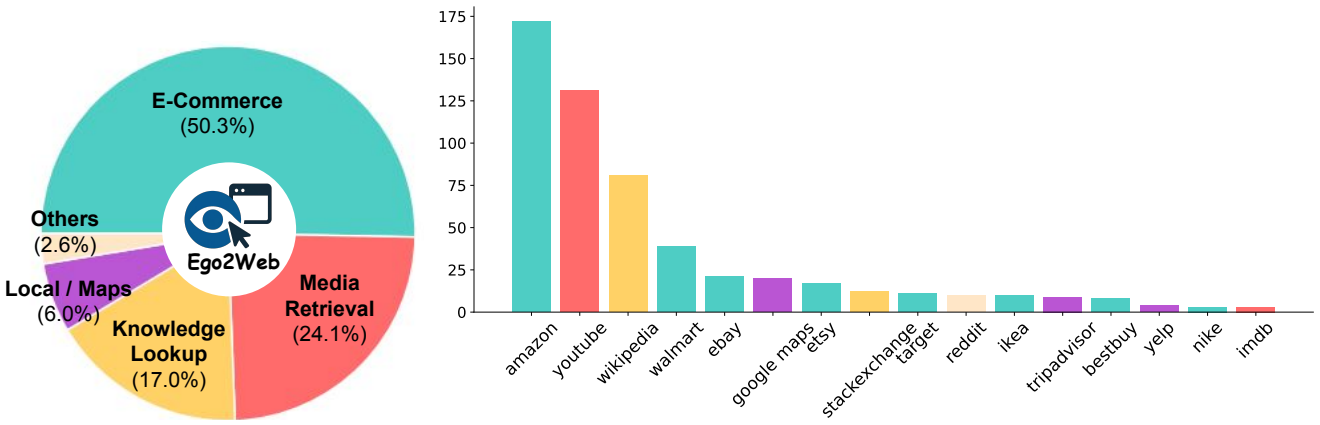
Video Captioning Prompt

You are a helpful assistant for video understanding. Your task is to carefully analyze the given video and provide a detailed description of its content.
Please describe both:

- **Global content**: overall scene, events, and actions
- **Local objects**: important objects, their attributes, and interactions

Domains / Agents	Claude 3.7	Claude 4.5	GPT 5.4	SeeAct	BU-GPT-4.1	BU-Genimi-3-Flash	Avg. SR
E-Commerce	13.0	18.2	14.3	19.5	26.9	**38.2**	21.7
Media Retrieval	19.6	26.5	29.5	24.2	30.3	**50.7**	30.1
Knowledge Lookup	33.6	45.6	39.1	43.4	63.0	**75.0**	50.0
Local / Maps	6.4	12.9	29.0	19.3	22.5	**48.3**	23.1
Others	0.0	6.6	6.6	20.0	**40.0**	13.3	14.4
Total	17.8	24.8	23.6	25.2	34.6	**48.2**	29.0

Table 5. Fine-grained Success Rate (SR) per task domain across different models, evaluated by `Ego2WebJudge` with Gemini-2.5 Pro.



(a) Task type distribution of `Ego2Web`.

(b) Distribution of website domains in `Ego2Web`.

Figure 5. **(a)** Task type distribution of `Ego2Web`, showing the high-level task category composition across major web platforms. **(b)** Fine-grained website domain distribution in Ego2Web (we omit websites whose count<4 for visualization), highlighting coverage across e-commerce, media retrieval, knowledge bases, local/map services and others (indicated by color coding).

Output Format. Your output must be a valid JSON object with the following schema:

```
{
    "video description":
    "a detailed description
    of the video content",
    "objects": {
        "object_name_1":
        "detailed
        description of
        object_name_1",
        "object_name_2":
        "detailed
        description of
        object_name_2"
    }
}
```

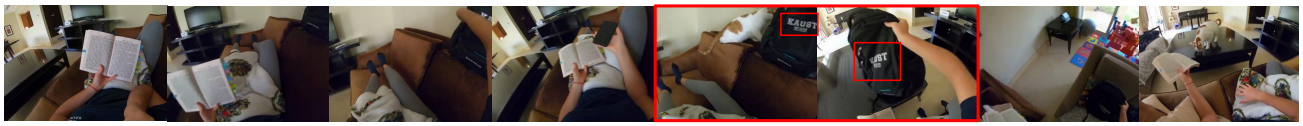
Important Notes.

- The `"video description"` should include key actions, temporal progression, and scene context.

Category	#Tasks	Websites
E-Commerce	230	amazon, ebay, walmart, apple, adidas, etsy, target, bestbuy, ikea, nike
Media Retrieval	132	youtube, imdb, bilibili
Knowledge Lookup	92	wikipedia, stackexchange
Local / Maps	31	google maps, yelp, tripadvisor
Others	15	linkedin, google calendar, reddit, nytimes, quora, booking

Table 6. Mapping from task categories to representative websites in `Ego2Web`, along with the number of tasks per category.

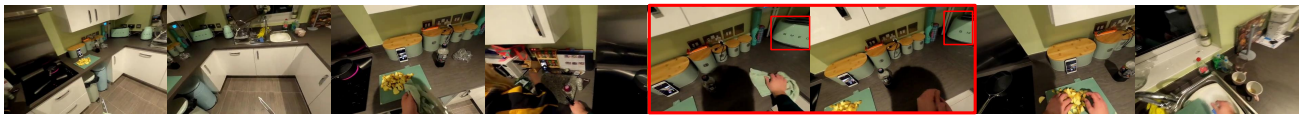
Egocentric Video



Web Task Instruction

On youtube.com, find an official campus tour or overview video from the university's official channel that matches **the name on the backpack**, and report the video's title and upload year.

Egocentric Video



Web Task Instruction

Find **the same color toaster seen in the video** on both amazon.com and ebay.com, and determine which site lists the lower price.

Egocentric Video

Web Task Instruction

On amazon.com, find **the second boxed snack bar that the shopper picks up** in the video and open its product page. Make sure the listing matches the exact flavor wording shown on the box. Then report the calories per bar and sugars per bar from the product listing.

Figure 6. Some examples of Ego2Web, including sampled frames of an egocentric video and paired web task instructions. We annotate essential video perception (e.g. the mint toaster with brand name and the black backpack in the middle of the video) with red boxes.

- Each object entry should include distinguishing attributes (e.g., color, shape, brand, state, interaction).
- If there are no clear objects, set:

```
"objects": {}
```

- Ensure the output is strictly valid JSON with no extra text.

Web Task Instruction Generation Prompt

User

You are given an egocentric video and its captions.
Detailed Video Description (every 5 seconds):
[Video Metainfo]

Step 1: Suitability Assessment

Determine if this video is suitable for designing a **video-grounded web task**.

A video is suitable if:

- It contains **objects, activities, or scenes** that naturally link to a **web-based information or action need**.
- The task requires **visual information** from the video.
- Visual cues are **clear and distinctive**.
- The task cannot be solved using captions alone.

If unsuitable, output '"suitable": false' and briefly explain why.

Step 2: Visual Anchor Extraction

If suitable, identify 1–2 visual anchors. For each:

- name guess
- 3–5 visual cues
- timestamps
- why video dependent

Step 3: Task Instruction Generation

For **one** selected anchor, design 1–3 task instructions:

- Must depend on ≥ 2 visual cues
- Must include a "Must Match" list
- Phrased as natural web instructions
- Clear, verifiable goal
- Include "why video dependent"

Restrict domains to:
["amazon.com","ebay.com","walmart.com", "aliexpress.com","etsy.com","ikea.com","nike.com", "adidas.com","apple.com","bestbuy.com","target.com", "google.com/maps","tripadvisor.com", "booking.com","airbnb.com", "expedia.com","openstreetmap.org", "wikipedia.org","reddit.com", "quora.com","stackexchange.com", "cnn.com","nytimes.com", "youtube.com","vimeo.com","imdb.com", "tiktok.com","bilibili.com", "yelp.com","weather.com","bbc.com", "docs.google.com", "calendar.google.com","notion.so","linkedin.com", "x.com","instagram.com","facebook.com"]

Output Format (strict JSON):

```
{
  "suitable": true/false,
  "reason if not suitable":
  "<short reason>",
  "tasks": [
    {
      "difficulty": 1,
      "instruction": "...",
      "must match": ["cue1", "cue2"],
      "timestamps": ["00:10-00:20"],
      "allowed domains": ["youtube.com",
      "amazon.com",
      "wikipedia.org"],
      "why_video_dependent": "..."
    },
    {
      "difficulty": 2,
      "instruction": "...",
      "must match": ["cue1","cue2",
      "cue3"],
      "timestamps": ["00:15-00:25"],
      "allowed domains": [
      "google.com/maps",
      "tripadvisor.com",
      "booking.com"
      ],
      "why_video_dependent": "..."
    }
  ]
}
```

Evaluation Prompt of `Ego2WebJudge`. We provide detailed prompts for the prompt for the proposed automatic evaluation framework, `Ego2WebJudge`, as shown in the following table.

`Ego2WebJudge` System Prompt

You are an expert evaluator for the Ego2Web benchmark. Your job is to determine whether a web navigation agent successfully completed a web task that is grounded in egocentric video evidence.
You are given:
- Egocentric video evidence (provided as sampled keyframes)
- The **task instruction**
- The agent's **action history**
- **Key points** for task completion
- **Potentially important webpage snapshots** from the agent's trajectory with explanations

Your goal is to determine whether the agent successfully completed the task **while correctly grounding the result in the egocentric video evidence**.

Strict Evaluation Principle. This benchmark requires **strict visual grounding**. A task should only be marked as *success* when there is **clear, direct, and consistent visual evidence** that the web result matches the objects, events, or actions shown in the egocentric video.
If there is **any uncertainty, ambiguity, mismatch, or missing visual evidence**, the task must be marked as *failure*.
Do **not** assume correctness based on:
- the agent's textual claim
- webpage titles
- search queries
- approximate or loosely related matches

False positives are worse than false negatives. When in doubt, mark the task as failure.

Important Evaluation Criteria.
1. **Filter correctness**. If filters are required, they must be correctly applied and visibly reflected in results. Missing selection, confirmation, or effect leads to failure.
2. **Proper use of filtering/sorting**. Constraints such as "best", "highest", "cheapest", "latest", "lowest", "closest", "highest-rated", "largest", and "newest" must be handled through actual filtering or sorting functions.
3. **Exact numeric constraints**. Ranges for price, year, beds, bathrooms, rating, etc., must match exactly. Any deviation results in failure. Examples:

- Requirement <$50 → Applied <$25 → failure
- Requirement $1500–$2500 → Applied $2000–$2500 → failure
- Requirement $25–$200 → Applied $0–$200 → failure
- Requirement 2004–2012 → Applied 2001–2012 → failure
- Requirement exactly 2 beds → Applied "2+ beds" → failure

Ego Video Grounding Rules.

4. The webpage result must be **strictly grounded in ego video evidence**, including object identity, category, brand, color, quantity, state, and action cues.
5. **Apparent success is not sufficient**. Even if the webpage looks correct, the task fails if it does not match the ego video.
6. Carefully verify **fine-grained visual details**: object identity, brand, color, material, shape, text, number of items, spatial relations, and actions.
7. For object retrieval tasks, the result must correspond to the **same object or correct category** supported by the video evidence.
8. For media retrieval tasks, the result must match the **same real-world event or action**.
9. If evidence conflicts, **prioritize visual grounding** over textual claims.
10. **Near or partial matches are considered failure** when precise identification is required.

Common Failure Cases. Failure includes cases such as:

- Product mismatch (e.g., different brand, color, or type than in the video)
- Similar-looking but incorrect objects
- Related but different events or actions in retrieved media
- Titles suggesting correctness but screenshots contradicting it
- Insufficient visual evidence
- Claims not supported by video or webpage evidence
- Partial matches that do not fully satisfy the task

Required Output Format. Your response must contain exactly two lines:

```
Thoughts: <reasoning based on
key points, webpage evidence,
and ego video evidence>
Status: success or failure
```

Do not output anything beyond these two lines.

Example of Video Caption. For agents, including Claude 3.7, Claude 4.5 and GPT-5.4, who can not access raw video input, we convert video into detailed video captions with timestamp via strong MLLM (Gemini-3.1-Pro). We show both video example and video caption example as follows:

Example of Detailed Egocentric Video Caption by Gemini-3.1-Pro

[00:00 - 00:05] The camera pans left to right across a glass-door refrigerator filled with various bottled and canned beverages, including sports drinks and sodas. The text 'Cold Water' is visible at the top of the cooler.
[00:05 - 00:10] The camera continues panning across the refrigerator shelves, revealing rows of brightly colored energy drinks and sparkling water cans. The camera movement is slightly blurry.
[00:10 - 00:15] A person's hand reaches into the refrigerator and grabs a blue can of Nos energy drink. The camera then quickly pans to a piece of paper being held over a red basket.
[00:15 - 00:20] A person holds a paper shopping list while pointing at it with a finger. A red shopping basket is visible below. The camera briefly shows the store's interior, including a pastry display case.
[00:20 - 00:25] The camera focuses closely on the paper, revealing a printed shopping list with items like 'Brownies', 'Beer', 'Something for non-beer drinkers', and 'Energy drinks'. A person is visible in the background.
[00:25 - 00:30] The camera quickly pans away from the list, showing a hot food display case and then sweeping across the store to show shelves stocked with snacks.
[00:30 - 00:35] The camera moves around the store, briefly showing a woman behind the checkout counter and then focusing on a multi-tiered display case filled with pastries.
[00:35 - 00:40] The camera pans to the checkout counter where a red shopping basket is placed. A woman with long dark wavy hair stands behind the counter, looking down. A sign reads 'PICK 5 FOR 8.00'.
[00:40 - 00:45] The woman behind the counter reaches for items near the register. The camera is slightly shaky, capturing the counter area and a small portion of the store's background.

[00:45 - 00:50] The camera pans rapidly from the checkout counter to the store's snack shelves and back, creating a blurred view of the store's interior layout.
[00:50 - 00:55] The camera focuses on the checkout counter. The woman with long dark wavy hair is seen working at the register, preparing to scan items from the red basket.
[00:55 - 01:00] The woman reaches into the red shopping basket and picks up a blue can, appearing to be Red Bull. She holds it up to scan it at the cash register.
[01:00 - 01:05] The woman scans the blue can, places it aside, and then picks up a green plastic bottle to scan it. She operates the cash register terminal with her other hand.
[01:05 - 01:10] She continues scanning items, picking up a pinkish bottle and then another tall green bottle. She efficiently processes the items from the customer's basket.
[01:10 - 01:15] The camera turns completely around to show the store's seating area, featuring a small table with chairs, two glass-door beverage refrigerators, and snack shelves.
[01:15 - 01:20] The camera pans across the store's interior, showing shelves stocked with various packaged snacks, and then sweeps back towards the checkout counter.
[01:20 - 01:25] The camera refocuses on the woman at the checkout counter. She is actively pressing buttons on the cash register. A 'HOUSE BLEND' coffee sign is visible in the background.
[01:25 - 01:30] The woman picks up a red packaged item from the counter, scans it, and places it into a brown paper shopping bag.
[01:30 - 01:35] She continues bagging the customer's purchases, placing items into the brown paper bag on the counter. A 'PICK 5 FOR 8.00' sign sits on the counter in front of her.
[01:35 - 01:40] The camera pans left to show a shelf display filled with various boxes and packages of snacks, then slowly moves back toward the counter area.
[01:40 - 01:45] The camera focuses closely on a tiered display rack stocked with different boxes of candies and small snacks, panning slowly across the items.
[01:45 - 01:50] The camera moves back to the checkout counter, showing the woman continuing to process the transaction. She handles more items and interacts with the cash register.
[01:50 - 01:55] The woman picks up a small rectangular packaged snack, scans it, and carefully places it into the brown paper bag along with the other items.
[01:55 - 02:00] She continues to place items into the brown paper bag. The camera then pans right, showing the hot food display case and part of the store's entrance.
[02:00 - 02:05] The camera pans to show a corner of the store serving as a small break area, containing a table, two black chairs, and recycling bins against the wall.
[02:05 - 02:10] The camera sweeps back toward the checkout counter, showing the woman still standing by the register, completing the bagging process.
[02:10 - 02:15] The woman is seen at the counter, handling the brown paper bag and arranging the items inside. She looks up briefly.
[02:15 - 02:20] She picks up another item from the counter, scans it, and places it into the brown paper bag. The 'PICK 5 FOR 8.00' promotional sign is clearly visible.
[02:20 - 02:25] The woman operates the cash register keyboard, pressing buttons to finalize the transaction. A green bottle stands on the counter beside the paper bag.
[02:25 - 02:30] She continues typing on the cash register keyboard, looking down at the screen. The filled brown paper bag sits ready on the counter.
[02:30 - 02:35] She picks up a yellow bag and a red package, scans them, and places them into the brown paper bag, continuing to pack the customer's items.
[02:35 - 02:40] She holds a red snack package in one hand while pressing buttons on the cash register with the other, confirming the final items.
[02:40 - 02:45] She looks at the items and interacts with the cash register one last time, seemingly finishing ringing up the customer's order.
[02:45 - 02:50] The camera tilts down to reveal a wire display rack attached to the front of the counter, fully stocked with various candies like M&Ms, Mentos, and Twix.
[02:50 - 02:55] The camera pans slowly across the colorful candy display. A customer's hand enters the frame holding several US dollar bills, preparing to pay.
[02:55 - 03:00] A close-up shows the customer's hand holding a fan of US dollar bills over the candy display, ready to hand the cash to the cashier.

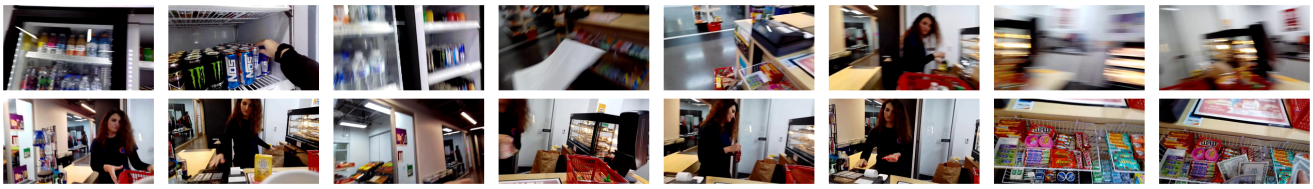


Figure 7. Example of egocentric video in Ego2Web, the detailed captions generated by Gemini-3.1-pro are listed above.