Active Intelligence in Video Avatars via Closed-loop World Modeling

Supplementary Material

1. Supplementary Material

This supplementary material provides comprehensive details to support the main paper. It is organized as follows: Section A details the construction of the L-IVA benchmark; Section B elaborates on the evaluation protocols and metrics; Section C presents additional qualitative visualizations comparing different methods; Section D lists the specific prompts utilized in the ORCA framework; Section E discuss the failure case and limitations.

2. Benchmark Construction Details

To ensure diversity in visual complexity and environmental dynamics, we construct the L-IVA benchmark through a hybrid pipeline combining real-world photography and AI-synthesized imagery.

Real-world Data Curation. We source high-quality real-world images from Pexels to serve as initial observations. The selection process is strictly guided by *scene affordance*: we filter for images containing distinct, interactive objects capable of supporting multi-step physical manipulations necessary to achieve high-level goals. For each selected scene, we manually define a high-level intention (I). To generate ground-truth annotations efficiently, we leverage Gemini-2.5-Pro. Given the image and the intention, the model generates a structured set of metadata, including sequential subgoals, detailed object descriptions, and reference action prompts. Each data sample is stored as a pair consisting of the initial image and a corresponding YAML file containing these hierarchical annotations (see Figure 1(c)).

Synthetic Data Generation. For synthetic scenarios, we utilize Nanobanana to create controlled environments with specific object configurations. Unlike the real-world pipeline, we adopt a *goal-first approach*: we first design a high-level intention and ensure that all requisite object interactions are logically solvable within a single scene. Based on these requirements, we craft detailed text prompts to generate the initial scene image. This "design-then-generate" strategy ensures precise alignment between the visual assets (objects in the scene) and the task requirements, guaranteeing that the generated environments inherently support the intended interaction sequence.

3. Evaluation Protocols

To systematically assess the capabilities of active video avatars, we employ a hybrid evaluation strategy combining VLM-based automated metrics and human judgment. All automated evaluations utilize `Gemini-2.5-Flash` due to its strong multimodal reasoning and long-context video understanding capabilities. We use user study to measure the physical score and TSR, as current VLMs struggle with subtle physical inconsistencies and long-horizon causal reasoning. The webpage of the user study is shown in Figure 5. Our user study involves 8 human evaluators. We constructed 130 comparative evaluation sets sampled across different scenarios. Since each set includes videos from 4 methods (ORCA + 3 baselines), this results in a total of 520 evaluated videos, ensuring statistical reliability. The evaluation focuses on three key dimensions:

Task Success Rate (TSR). TSR measures the agent's ability to complete the high-level intention through multi-step subgoals. Unlike single-clip generation, our task requires causal completion of a sequence of actions. Human evaluators are presented with the high-level intention I, the ground-truth subgoal list $\{g_1, ..., g_N\}$, and the generated video sequence V. For each sample, evaluators count the number of *successfully completed subgoals* based on visual evidence. The final TSR score is calculated as the ratio of completed subgoals to the total number of subgoals:

$$\text{TSR} = \frac{1}{K} \sum_{k=1}^{K} \frac{N_{\text{completed}}^{(k)}}{N_{\text{total}}^{(k)}} \tag{1}$$

where $N_{\text{completed}}^{(k)}$ is the number of subgoals successfully executed in the k-th test case.

Action Fidelity Score (AFS). AFS measures the semantic alignment between the planned action command a_t and the executed video clip v_t. As detailed in **Table 1**, this is a binary classification task $(0/1)$. The VLM verifies if the *Core Action*, *Key Objects*, and *Directional Consistency* in the video match the text caption. It is designed to be tolerant of minor visual artifacts but strict on semantic errors (e.g., performing "cut" instead of "peel").

Physical Plausibility Score (PPS). This metric evaluates the physical consistency of the generated world, specifically targeting I2V-specific hallucinations. Human evaluators score each video sequence on a **1-5 Likert scale**:

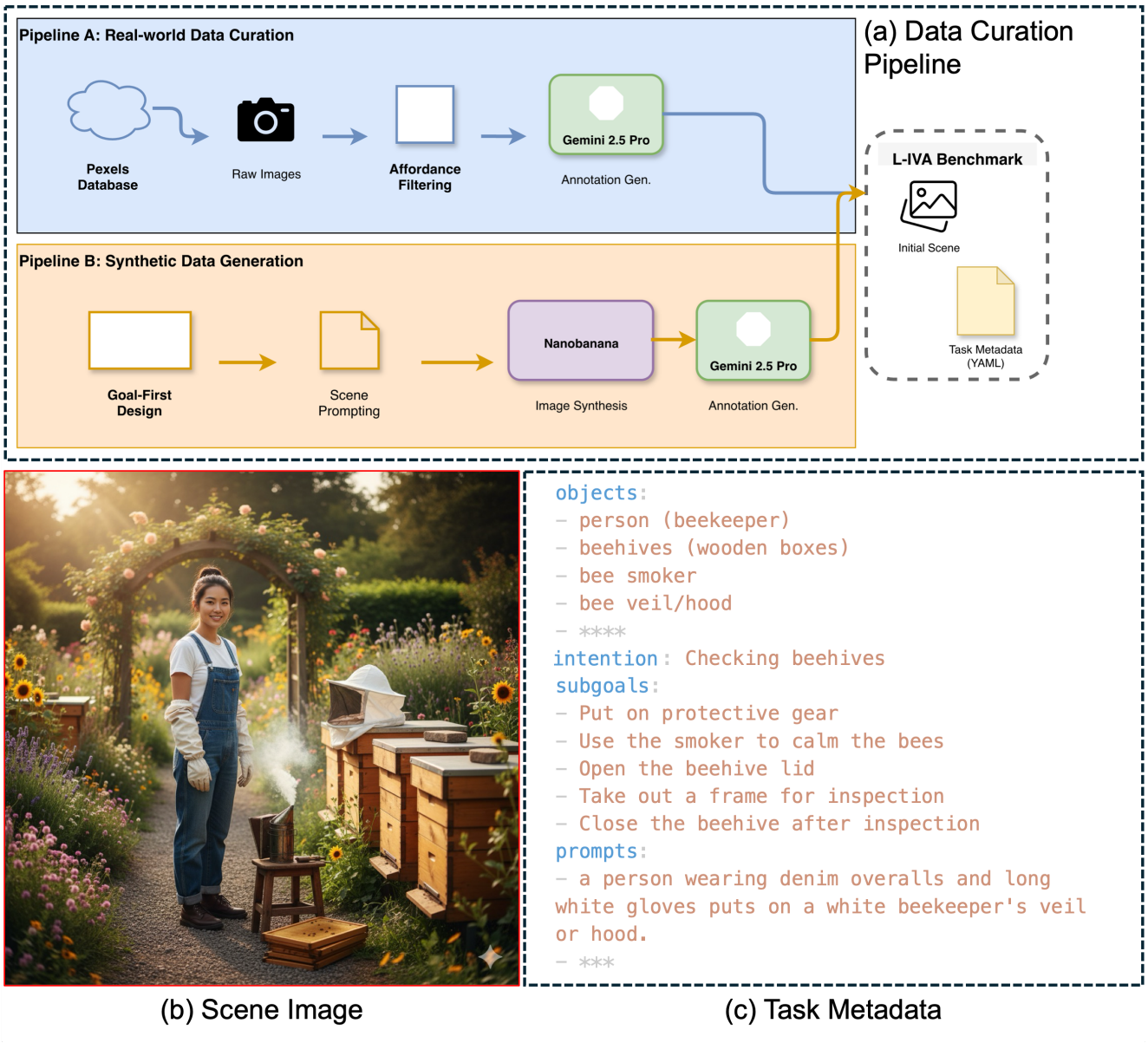

Figure 1. **Overview of the L-IVA Benchmark Construction Pipeline.** (a) Our data curation process employs a hybrid strategy: **Pipeline A** sources real-world images from Pexels, filtered by scene affordance and annotated via Gemini-2.5-Pro. **Pipeline B** utilizes a goal-first design for synthetic data, where scenes are generated by Nanobanana to strictly align with intended interactions. (b) A representative scene image (e.g., "Checking beehives") from the benchmark. (c) The corresponding structured metadata (YAML), including object inventory, high-level intention, subgoals, and reference prompts.

- **5 (Perfect):** Perfect physical interaction with realistic gravity, collision, and contact points.
- **4 (Good):** Physics generally correct; minor artifacts do not impede comprehension.
- **3 (Fair):** Noticeable floating or clipping, but the action logic remains coherent.
- **2 (Poor):** Severe physical violations (e.g., object teleportation, interpenetration).
- **1 (Fail):** Complete breakdown; failure to adhere to basic physical laws.

Human Preference Ranking Protocol. To evaluate holistic performance, we employ Best-Worst Scaling (BWS), which is shown to be more statistically robust than simple ranking. For each query, annotators are presented with the initial scene, the intention, and anonymized videos

generated by the comparing methods side-by-side. Annotators are instructed to select the **Best** and **Worst** models based on a hierarchical criterion: first prioritizing *Task Completion* (whether the intention was fulfilled), followed by *Logical Coherence* (smoothness of the action plan), and finally *Visual Aesthetics*. The BWS score for each method is computed as:

$$\text{BWS Score} = \frac{\#\text{Best} - \#\text{Worst}}{\#\text{Total Cases}} \times 100\% \tag{2}$$

This score ranges from -100% to +100%, where a positive score indicates that the method was chosen as "Best" more often than "Worst".

4. Prompt Details for ORCA

In this section, we provide the full system prompts used in the ORCA framework. These prompts are designed to be model-agnostic but are optimized for Gemini-2.5-Flash.

- **Figure 3**(a): The *Initialization Prompt* used by System 2 to parse the initial scene and decompose the high-level intention into a structured plan.
- **Figure 3**(b): The *Observation Prompt* for updating the belief state based on generated video clips.
- **Figure 4**(a): The *Thinking Prompt* (System 2) for strategic reasoning, discrepancy detection, and next-step planning.
- **Figure 4**(b): The *Action Grounding Prompt* (System 1) for translating abstract plans into high-fidelity, I2V-compatible captions.
- **Figure 4**(c): The *Reflection Prompt* for verifying action execution and triggering error correction.

5. Discussion on Failure Cases and Limitations

While ORCA demonstrates superior performance in long-horizon task execution compared to open-loop baselines, its capabilities are inevitably bounded by the underlying foundation models (VLM and I2V). We categorize the observed failure cases into two dimensions: *Perceptual Bottlenecks* (VLM-side) and *Generative Constraints* (I2V-side). It is important to note that these failures stem primarily from the intrinsic limitations of current pre-trained models rather than the algorithmic design of the ORCA framework.

VLM-Centric Failures: Perception and Depth Ambiguity. The reliability of ORCA's *Reflect* and *Observe* modules depends on the VLM's visual grounding ability. We observe two specific issues:

- **Temporal Information Loss due to Sampling:** To manage context length, ORCA feeds sampled frames (e.g., 5 frames) to the VLM. This discrete sampling can cause *Temporal Aliasing*, where critical but fleeting glitches (e.g., an object flickering out of existence for just 2 frames) fall between sampled frames. Consequently, the VLM may generate a "False Positive" judgment, accepting a flawed video.
- **Lack of 3D Spatial Awareness:** Current VLMs operate in 2D pixel space and often struggle with depth perception. As illustrated in Figure 2(b), the VLM may misinterpret a background object as being within the avatar's reach, leading to geometrically impossible instructions (e.g., "pick up the distant cup") that the I2V model cannot execute realistically.

I2V-Centric Failures: Control and Consistency. Even when ORCA generates perfect instructions, the generative backbone (I2V model) acts as a physical execution bottleneck:

- **Instruction Following:** In handling complex, fine-grained manipulations, the I2V model often exhibits strong prior biases and fails to adhere to the prompt. Although ORCA's *Reflect* module correctly rejects these failures and triggers retries (up to N_{retry} times), the I2V model may persistently fail to generate the correct physics, leading to an eventual task termination.
- **Object Permanence and Hallucination:** Generative models inherently struggle with long-term object permanence. As shown in Figure 2(d), sudden object disappearance or the spontaneous hallucination of new objects can occur. While ORCA's state tracking attempts to catch these errors, severe hallucinations can sometimes corrupt the agent's belief state if they are too subtle for the VLM to detect immediately.

Conclusion. These limitations highlight that ORCA is a *framework for active intelligence*, currently operating on imperfect substrates. We posit that as VLM spatial reasoning and I2V controllability improve, ORCA's performance will scale accordingly without architectural changes.

Figure 2. **Qualitative Analysis of Failure Cases attributed to Foundation Model Limitations.** (a) **Temporal Information Loss:** Due to discrete frame sampling, the VLM misses the ”teleportation” artifact where the fertilizer bag instantaneously appears in the hand (red arrow), falsely accepting it as a valid pickup. (b) **Lack of 3D Spatial Awareness:** The VLM misinterprets the depth of the scene, instructing the avatar to fetch a watering can that is actually in the distant background, resulting in an unnatural reaching motion. (c) **Weak Instruction Following:** For fine-grained tasks like ”light the alcohol lamp,” the I2V model consistently fails to execute the interaction despite ORCA triggering multiple retries (separated by dashed lines). (d) **Object Disappearance:** A clear example of generative instability where a key object (the water bottle) vanishes mid-clip (red box) despite no interaction occurring.

Table 1. Full Prompt for Action Fidelity Score (AFS) Evaluation

Prompt Content

```
# ROLE: Action Fidelity Evaluator
You are an expert evaluator assessing alignment between a video clip and its action caption.

## EVALUATION CRITERIA
ALIGNED (af = 1) if:
1. Core Action Match: Primary verb (e.g., "pick up") is clearly visible.
2. Object Correctness: Key objects match the caption.
3. Spatial/Directional Consistency: Motion direction and location match roughly.

NOT ALIGNED (af = 0) if:
1. Wrong Action: Action shown is fundamentally different (e.g., "cut" vs "stir").
2. Wrong Object: Object manipulated is incorrect.
3. Missing Action: Described action is absent.
4. Impossible Outcome: Visual contradicts caption's physics.

## TOLERANCE GUIDELINES
- ACCEPT: Visual quality issues (blur), Incomplete visibility (partially off-screen), Semantic equivalence ("bowl" vs
"container"), Timing flexibility.
- REJECT: Contradictory actions, Wrong action type, Missing critical steps.

## OUTPUT FORMAT
Provide reasoning in <thinking>, then output JSON:
{
  "af": 0 or 1,
  "reason": "Brief explanation..."
}
```

```
# ROLE: Cognitive System - Initialization Module
# MISSION: Analyze scene and decompose intention into a feasible plan.

# CONTEXT
- High-Level Intention: "{intention}"
- Visual Input: Initial scene image.

# YOUR TASK
1. SCENE ANALYSIS: Identify ALL interactive objects (ID, description,
state) and Avatar state.
2. PLAN GENERATION: Create a `task_checklist` strictly using the
identified inventory.

# CORE PLANNING PRINCIPLES
1. Inventory Principle: Only use what you see.
2. Common Sense Inference: Can assume internal states (e.g., kettle has
water) but NO invisible objects.
3. Logical Flow: Sequential order.
4. Simplicity: High-level sub-goals (e.g., "Grind coffee"), not atomic moves.
5. No Magic: No teleportation.
6. Verifiable Outcome: Each step must have a visible state change.

# OUTPUT FORMAT
<thinking>reasoning</thinking>
```json
{
"scene_description": "...",
"interactive_objects": { "id": {"description": "...", "state": "..."} },
"avatar_state": { "right_hand": "...", "left_hand": "..." },
"task_checklist": [ {"goal": "...", "status": "pending"} ]
}
```

(a) Prompt for Initialization Module
```

```
# ROLE: Perceptual System - State Estimator
# MISSION: Update belief state based on new visual evidence.

# INPUTS: Previous Belief + New Video Frames.

# TASK: DETECT CHANGES AND UPDATE STATE
1. Compare visual evidence against previous belief.
2. Identify all changes and construct a complete, updated JSON.

# JSON SCHEMA (Mandatory)
{
"scene_description": "...",
"interactive_objects": { "id": { "description": "...", "state": "..." } },
"avatar_state": { "right_hand": "...", "left_hand": "..." }
}

# RULES
- ALL KEYS REQUIRED. Return empty objects/null if no info.
- DO NOT omit `avatar_state`.
```

(b) Prompt for Observer

Figure 3. Prompt for Initialization module and observer

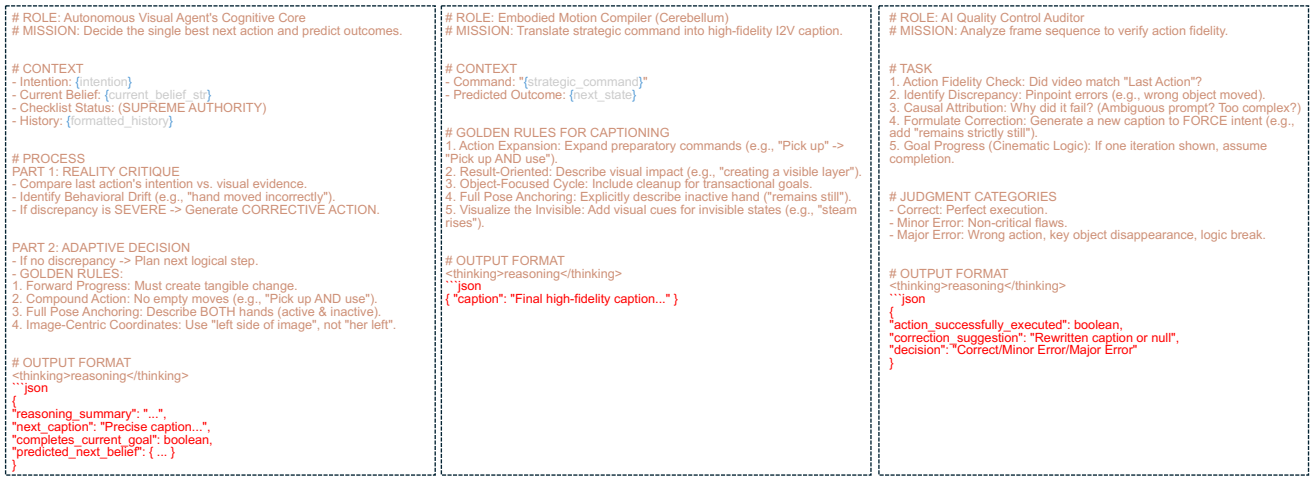

Figure 4. Prompt for Thinking, System-1 and Reflector

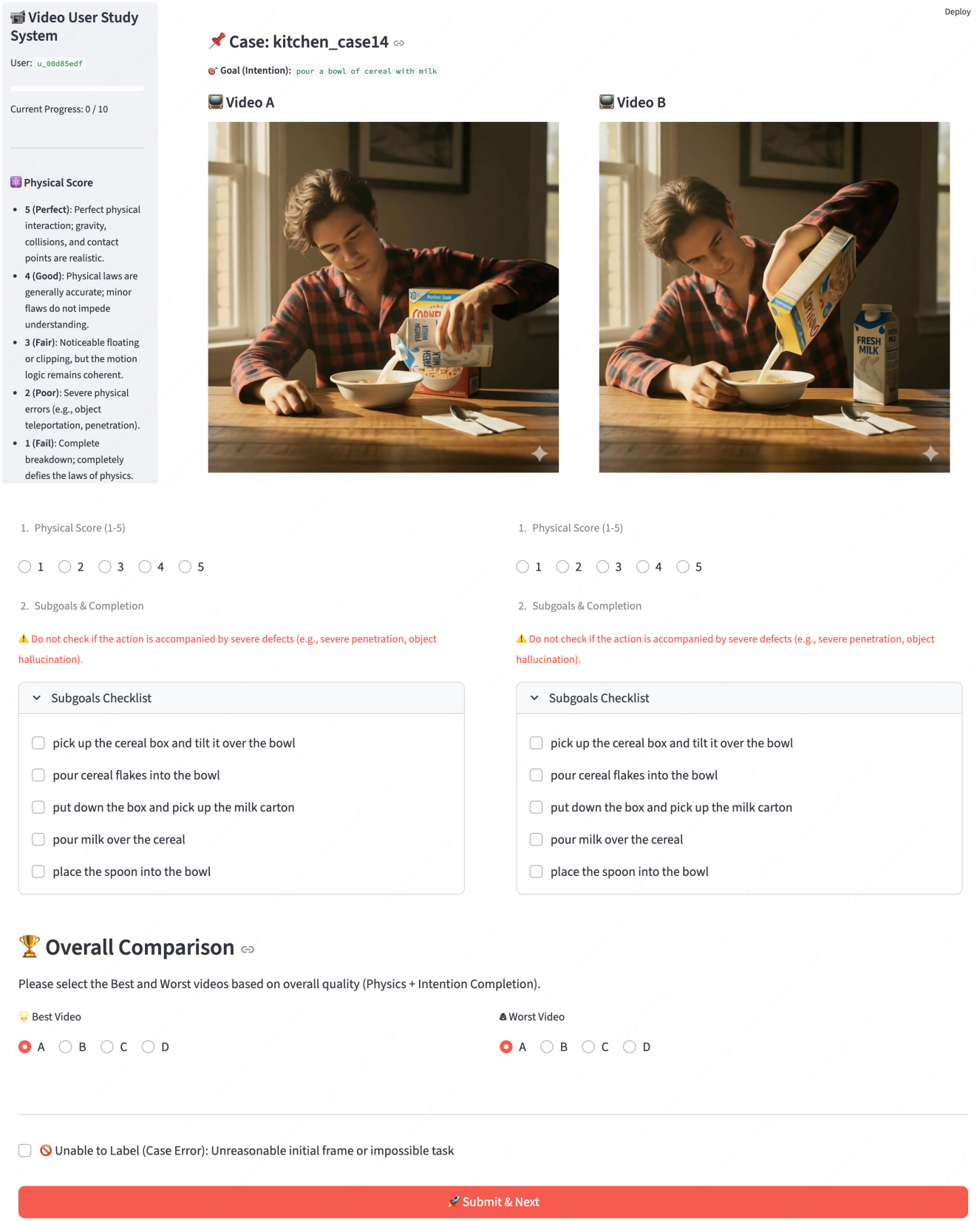

Figure 5. The user interface for human evaluation. For each test case, annotators are presented with a high-level goal (Intention) and generated videos from four anonymized methods (Video A and Video B are shown here; C and D are omitted for brevity). Evaluators are asked to assess each video based on two criteria: (1) a Physical Score (1-5 Likert scale) regarding simulation stability, and (2) a Subgoals Checklist to verify task completion. Finally, they select the best and worst videos based on overall quality.