BEHAVIOR TURING TEST: EVALUATING HUMAN-LIKENESS OF HUMANOID BEHAVIORS IN CONTEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in humanoid robotics have enabled increasingly complex behaviors in real and simulated environments. However, existing evaluation remains largely motion-centric, focusing on task success, tracking accuracy, or motion-level human-likeness, while leaving a key question unexplored: *would a human behave this way under the same task and scene?* To address this, we introduce the **Behavior Turing Test (BTT)**, which evaluates human-likeness at behavior level under explicit task and scene context. We construct **HuCo**, a dataset of 1,280 behavior clips spanning real humans, digital humans, real humanoids, and simulated humanoids, each paired with a Task Goal and Scene Description. We introduce **Context-SMPL** to suppress embodiment-specific appearance cues while preserving motion and scene context, and collect human judgments of contextual human-likeness. Our analysis shows that some humanoid behaviors can pass the Motion Turing Test yet fail BTT once task and scene context are considered, and that a clear human-humanoid gap persists under contextual evaluation. Building on HuCo, we establish **HuCoBench** for BTT and construct task and scene counterfactual subsets to evaluate context sensitivity. We further introduce **CoBE**, a **Co**ntext-conditioned **B**ehavior **E**valuator that retrieves context-relevant behavioral evidence and achieves the strongest alignment with human judgments among evaluated methods. We will release dataset and code to support the community.

1 INTRODUCTION

Recent advances in humanoid robotics have expanded humanoid capabilities from basic locomotion to whole-body manipulation and complex task execution (Fu et al., 2024; He et al., 2024; Jiang et al., 2024b). Meanwhile, human motion generation and humanoid control are increasingly moving toward behaviors conditioned on tasks, objects, and surroundings (Araújo et al., 2023; Lin et al., 2026; Allshire et al., 2025; Xie et al., 2026). Yet evaluation remains largely *motion-centric*, focusing on naturalness, physical plausibility, and motion-text consistency (Meixner et al., 2024; Guo et al., 2022). Recently, the Motion Turing Test (MTT) (Li et al., 2026) asks whether an observed motion appears human-like. Yet a humanoid can move like a human while behaving inappropriate for the intended task or surrounding scene. This raises a broader question: *Would a human behave in this way under the same task and scene?* As illustrated in Fig. 1, answering this question requires evaluating the motion together with its task and scene context.

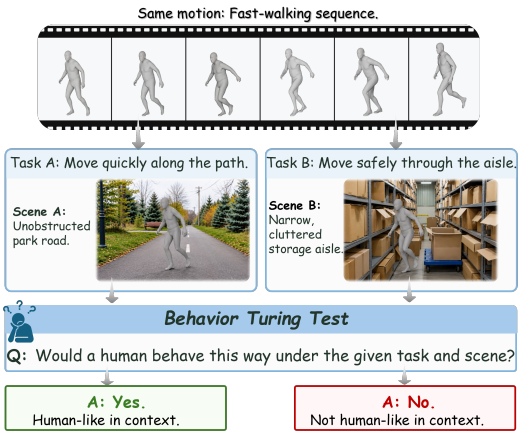

Figure 1: Motivation for **Behavior Turing Test.** The same human-like motion can lead to different behavior-level judgments under different context.

Motivated by this distinction, we introduce the **Behavior Turing Test** (BTT), a contextual human-likeness evaluation task. Given an observed behavior, its intended **Task Goal**, and **Scene Descrip-**

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tion, BTT assesses how closely the behavior matches natural human behavior under the same conditions. It evaluates both motion execution and whether the selected action is appropriate for the intended task and environment. Under this formulation, a behavior may pass MTT yet fail BTT when its motion appears human-like but its action is inconsistent with the task or scene.

To support BTT, we construct the **Hu**man-likeness of **Co**ntextual Behavior (**HuCo**) dataset, with **1,280 behavior clips** spanning real humans, digital humans, real humanoid robots, and simulated humanoids (Xie et al., 2026; Lu et al., 2025; NVIDIA, 2026b; Lin et al., 2026; NVIDIA, 2026a). Each clip is paired with an explicit **Task Goal** and behavior-relevant **Scene Description**. Since raw videos contain embodiment-specific appearance cues, while removing the scene discards essential context, we introduce a unified **Context-SMPL** representation. We remove the original subject using segmentation and video inpainting methods (Ren et al., 2024; Zhou et al., 2023), then re-render the recovered SMPL-X motion (Shen et al., 2024; Pavlakos et al., 2019) into the original scene. Context-SMPL suppresses embodiment-specific appearance cues while preserving motion and environmental context. All samples undergo manual verification and stage-wise quality control.

To establish human reference judgments for BTT, we collect **Contextual Human-Likeness Scores** on a 0-5 scale, measuring how closely a behavior matches how a human would naturally act under the same task and scene. Higher scores indicate greater human-likeness. We also collect **Motion Human-Likeness Scores** following the MTT protocol (Li et al., 2026). The two judgments are strongly correlated, yet contextual scores are consistently lower across all behavior domains. More importantly, some humanoid behaviors pass MTT but fail BTT once task and scene context are considered. This reveals a key distinction between *moving like a human* and *behaving like one in context*. We further construct **Task-Counterfactual** (Task-CF) and **Scene-Counterfactual** (Scene-CF) subsets by preserving the motion while changing the task or scene. These rating shifts provide controlled evidence that contextual human-likeness depends on more than motion alone.

Building on HuCo and its human annotations, we establish **HuCoBench**, a benchmark for BTT. Given a behavior video, Task Goal, and Scene Description, an evaluator predicts a 0-5 contextual human-likeness score aligned with human judgments. We evaluate three model families: general-purpose VLMs, human-supervised VLMs, and dedicated scoring models (Google DeepMind, 2026; Qwen Team, 2026e; Wang et al., 2025c; Qwen Team, 2026a). General-purpose VLMs show limited alignment, while human-supervised fine-tuning substantially improves performance. We further introduce the **Co**ntext-conditioned **B**ehavior **E**valuator (**CoBE**), a dedicated evaluator that uses task and scene context to retrieve relevant behavioral evidence from video features. CoBE achieves the strongest alignment with human judgments among the evaluated methods.

Overall, BTT extends human-likeness evaluation from motion to behavior under explicit task and scene context. HuCo and HuCoBench provide the data and evaluation framework for systematically studying this problem, while CoBE provides a context-aware computational evaluator. By making the gap between *moving like a human* and *behaving like one* measurable, BTT provides an important evaluation mechanism for advancing humanoid robots toward task and scene-appropriate behavior in complex real-world environments.

2 Related Work

2.1 Human and Humanoid Datasets

Human and humanoid datasets have progressively expanded from isolated motion to interaction with objects and scenes. Human motion datasets primarily capture body trajectories and action semantics from motion capture or video (Ionescu et al., 2014; Mahmood et al., 2019; Plappert et al., 2016; Punnakkal et al., 2021; Yan et al., 2024; Guo et al., 2022; Lin et al., 2023). Interaction datasets further incorporate body-object relations and scene context (Yan et al., 2023; Bhatnagar et al., 2022; Li et al., 2023; Lu et al., 2025; Hassan et al., 2019; Araújo et al., 2023; Wang et al., 2022; Jiang et al., 2024a). For humanoids, recent datasets and data pipelines provide motion references for imitation and whole-body control (Mao et al., 2025; Lee et al., 2025; He et al., 2024), with increasing coverage of loco-manipulation and scene interaction (Yang et al., 2026; Liu et al., 2024; Xie et al., 2026; Zhao et al., 2026). A few works jointly study human and humanoid motion for perceptual human-likeness or human-robot identification (Li et al., 2026; Kim et al., 2026). However, existing resources do not

Table 1: Comparison of HuCo with representative datasets. HuCo combines diverse human and humanoid behaviors with explicit task-scene context and human ratings.

Dataset	Human		Humanoid		Context		Human Evaluation		#Clips	#Hours
	Real	Digital	Real	Sim.	Scene Interact.	Task+Scene Text	Motion Human-Likeness	Contextual Human-Likeness		
AMASS (Mahmood et al., 2019)	✓	–	–	–	–	–	–	–	11.3K	40+
HumanoidRobotPose (Amini et al., 2021)	–	–	✓	–	–	–	–	–	0.023K	–
BEHAVE (Bhatnagar et al., 2022)	✓	–	–	–	✓	–	–	–	0.321K	–
Motion-X (Lin et al., 2023)	✓	–	–	–	–	–	–	–	81.1K	144.2
CIRCLE (Araújo et al., 2023)	✓	–	–	–	✓	–	–	–	7K+	10
LAFAN1 Retarget (lvhaidong, 2025)	–	–	–	✓	–	–	–	–	–	–
AMASS Retarget (Zhao, 2025)	–	–	–	✓	–	–	–	–	–	–
ViMoGen-228K (Lin et al., 2026)	✓	✓	–	–	✓	–	–	–	228.2K	369.4
GRAIL (Xie et al., 2026)	–	–	–	✓	✓	–	–	–	20K+	–
HHMotion (Li et al., 2026)	✓	–	✓	✓	–	–	✓	–	1.0K	1.39
HvH (Kim et al., 2026)	✓	–	✓	–	–	–	–	–	0.574k	2.8
HuCo (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	**1.28K**	**4.3**

jointly provide diverse human and humanoid behaviors, explicit task and scene context, and human judgments of contextual human-likeness. HuCo fills this gap, as summarized in Table 1.

2.2 Human-Likeness and Contextual Behavior Evaluation

Human motion evaluation traditionally focuses on generation quality, semantic consistency, and physical plausibility, using metrics such as joint errors, Fréchet Inception Distance (FID), motion-text alignment, and contact-based constraints (Guo et al., 2020; 2022; Tevet et al., 2023; Lin et al., 2026). Recent learned evaluators further incorporate human perceptual preferences to better align motion assessment with human judgments (Wang et al., 2025b). In embodied settings, evaluation often emphasizes task success, robustness, and physical interaction quality (Li et al., 2022; Sferrazza et al., 2024; Liu et al., 2024; Yu et al., 2026). Human-likeness has also emerged as an explicit evaluation target. The Motion Turing Test (Li et al., 2026) evaluates whether an observed motion appears human-like, while I-Robot (Kim et al., 2026) studies human-humanoid motion identification. Related work in LLMs evaluates whether generated behaviors resemble human behavioral patterns (Duan et al., 2024; Lei et al., 2026; Hathidara et al., 2026). These approaches do not explicitly evaluate whether an observed behavior is appropriate for its intended task and surrounding scene. As a result, a motion may appear human-like under motion-level evaluation while the corresponding behavior remains inconsistent with its context. BTT targets this distinction by evaluating human-likeness jointly over behavior, task, and scene.

3 HuCo: Human-Humanoid Contextual Behavior Dataset

We present **HuCo**, the dataset supporting the Behavior Turing Test. HuCo contains human and humanoid behaviors paired with explicit task and scene context and human judgments of contextual human-likeness. The complete data curation pipeline is provided in Fig. A1.

3.1 Dataset Overview and Data Sources

HuCo contains **1,280 behavior clips** evenly distributed across four domains: **real humans**, **digital humans**, **real humanoids**, and **simulated humanoids**. Each clip contains a clear task objective and meaningful interaction with the surrounding environment, and is paired with a **Task Goal** and a behavior-relevant **Scene Description**. All clips are converted into the unified Context-SMPL representation for human evaluation.

Real Humanoids. We collect publicly available demonstrations of recent humanoid systems from official websites, public video platforms, and released research demonstrations (Robotics, 2024b;a; Tesla, 2022). The initial collection contains 650 clips from 21 humanoid robot models. We retain behaviors with explicit task execution, meaningful scene and object interaction, and exclude isolated locomotion and context-free motions. **Simulated Humanoids.** Simulated humanoid behaviors are collected from GRAIL (Xie et al., 2026) and the NVIDIA PhysicalAI WorldModel Synthetic Em-

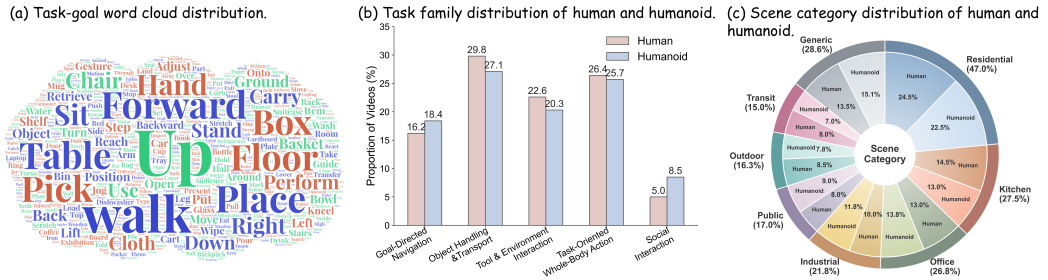

Figure 2: Statistics of HuCo. (a) Task-goal word cloud. (b) Task-family distributions across human and humanoid behaviors. (c) Scene-category distributions across human and humanoid behaviors.

bodied Robot Scenes dataset (NVIDIA, 2026b). These sources cover diverse navigation, and object-interaction, and task-execution behaviors across a range of simulated environments.

Real Humans. We select real-human behaviors from HUMOTO (Lu et al., 2025), BEHAVE (Bhatnagar et al., 2022), and the in-the-wild subset of ViMoGen-228K (Lin et al., 2026). We retain sequences with clear task intent and observable human-object or human-scene interaction, providing behavioral references under diverse real-world contexts. **Digital Humans.** Digital-human behaviors are collected from the NVIDIA PhysicalAI WorldModel Synthetic Digital Human Scenes dataset (NVIDIA, 2026a) and ViMoGen-228K (Lin et al., 2026). These samples extend HuCo with generated behaviors and environments beyond real-world recordings.

After rigorous filtering and manual verification, HuCo contains 320 clips from each domain, totaling **1,280 behavior clips**. Fig. 2 summarizes the task-goal vocabulary and the task and scene distributions across human and humanoid samples. We maintain comparable task and scene distributions between the two groups to reduce differences in behavioral complexity and scene composition. Detailed source composition and filtering procedures are provided in Appendix A.1-A.3.

3.2 Context-SMPL Video Construction

Behavior Turing Test requires a representation that suppresses embodiment-specific appearance cues while preserving the motion and scene context needed for behavioral judgment. Raw videos expose strong visual differences across humans and humanoids, which may influence human or model evaluations. We therefore convert all clips into a unified **Context-SMPL** representation, preserving the subject's motion and its spatial relation to the surrounding environment.

As illustrated in Fig. 3, Context-SMPL construction consists of three stages. We first segment and remove the original subject, then reconstruct the actor-free background using video inpainting method (Ren et al., 2024; Zhou et al., 2023). We next recover the subject's SMPL-X motion from the input video (Pavlakos et al., 2019; Shen et al., 2024). The recovered motion is finally re-rendered into the original scene while preserving its scene-relative position and trajectory. Stage-wise quality control and manual inspection are applied throughout the pipeline to remove samples with unreliable motion recovery, background reconstruction, or scene alignment. Implementation details and exclusion criteria are provided in Appendix B.

3.3 Dataset Annotation and Quality Control

Task and Scene Annotation. BTT evaluates behavior under an explicit task and scene context. We therefore annotate each clip with a **Task Goal**, describing the intended objective, and a **Scene Description**, capturing environmental information relevant to the observed behavior. We use Gemini 3.5 Flash (Google DeepMind, 2026) to generate initial annotations from raw videos sampled at 5 fps, with available source metadata used only as auxiliary references. The model is instructed to focus on visible behavioral and scene evidence, ignore embodiment-specific information, and omit task-irrelevant background details. All annotations are manually reviewed and corrected when necessary. The full annotation prompt is provided in Appendix C.1.

Human Annotation. We collect human judgments for all HuCo clips to establish the reference scores for BTT. Annotators view each Context-SMPL video together with its Task Goal and Scene

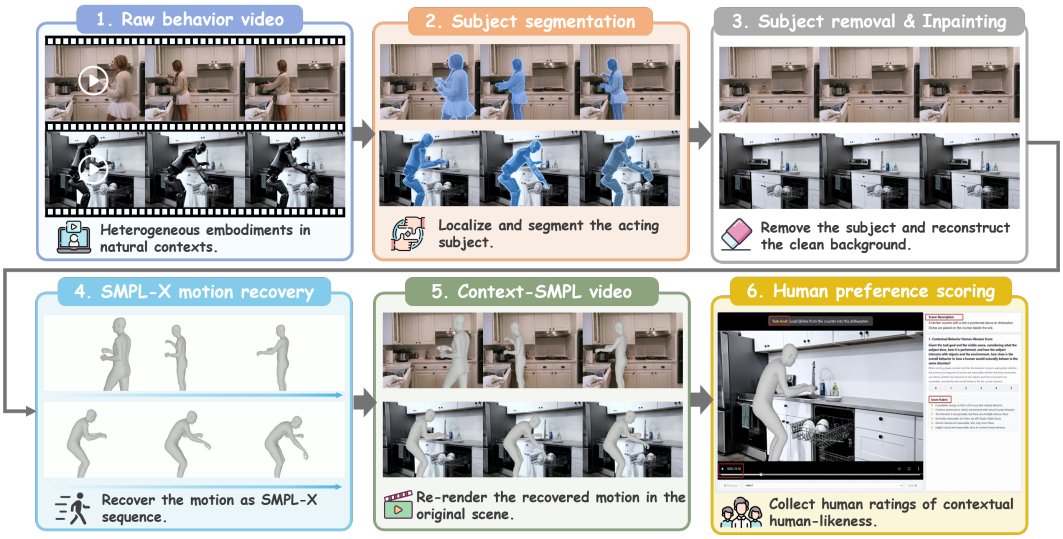

Figure 3: Context-SMPL construction and human evaluation pipeline. The original subject is removed, the recovered SMPL-X motion is re-rendered into the scene, and annotators evaluate the behavior conditioned on its Task Goal and Scene Description.

Description, as illustrated in Fig. 3. They rate how closely the observed behavior matches how a real human would naturally act under the same task and scene. Each sample receives a **Contextual Human-Likeness Score** from 0 to 5 (Likert, 1932), with higher scores indicating stronger agreement with natural human behavior. Samples are presented in randomized order, while the original embodiment type and data source are hidden from annotators. The scoring rubric and annotation interface are provided in Appendix C.2.

Annotation Quality Control and Background Bias. We assess inter-annotator agreement using ICC(2,N) (Shrout & Fleiss, 1979) and inspect inconsistent ratings during quality control. Since Context-SMPL preserves the original scene, we also test whether the background alone reveals the source domain. We construct a background-only subset by removing the acting subject and ask annotators to classify each background sample as originating from a human or humanoid source. Performance remains close to chance, indicating that the preserved background does not provide a reliable source-discriminative cue. Detailed reliability and background-bias results are reported in Tables A3 and A4, with protocols in Appendix C.3 and C.4.

3.4 Dataset Analysis

Contextual Human-Likeness across Domains. Fig. 4(b) summarizes contextual human-likeness across the four domains, together with representative behaviors examples in Fig. 4(a). Real humans achieve the highest **average score** (4.347), followed by digital humans (4.019), simulated humanoids (3.566), and real humanoids (2.905). Among humanoids, real humanoids contain a slightly larger proportion of samples in the $[3, 4]$ range, while simulated humanoids contain more samples in $[4, 5]$. Overall, both humanoid domains remain below the two human domains, revealing a persistent human-humanoid gap under explicit task and scene context.

Comparison with Motion Human-Likeness. To examine the relationship between contextual and motion-level human-likeness, we also collect **Motion Human-Likeness Scores** following the MTT protocol (Li et al., 2026). Per-domain comparisons are reported in Appendix D.1 (Table A5 and Fig. A7). Across all four domains, contextual and motion scores are strongly correlated (Spearman's $\rho = 0.831$–0.957), yet contextual scores are consistently lower. More importantly, using the average Real Human scores as human-reference criteria, we identify humanoid behaviors that **pass MTT but fail BTT**: their motion human-likeness exceeds the average Real Human MTT score, while their contextual human-likeness falls below the average Real Human BTT score (Appendix D.1). This directly supports the central distinction motivating BTT: moving like a human does not guarantee behaving like one under the same task and scene. A representative example is shown in Fig. 4(a).

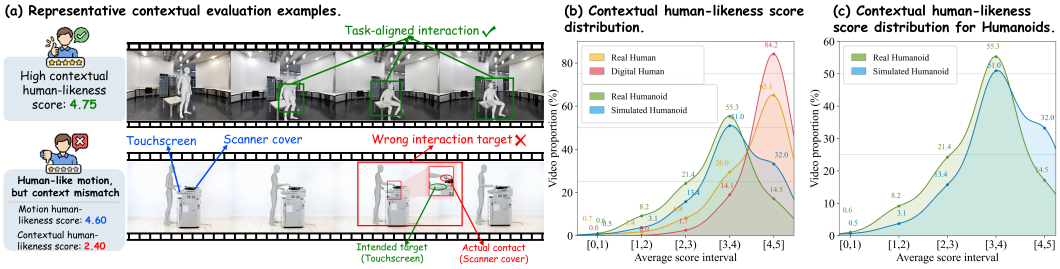

Figure 4: Contextual human-likeness analysis. (a) Representative examples, including a high motion human-likeness but low contextual human-likeness. (b) Score distributions across four behavior domains. (c) Score distributions for real and simulated humanoids.

4 HuCoBench: A Benchmark for the Behavior Turing Test

Building on HuCo and human annotations, we establish **HuCoBench**, a benchmark for evaluating how well computational evaluators align with human judgments of contextual human-likeness. We consider three evaluator families: general-purpose VLMs, human-supervised VLMs, and dedicated scoring models together with our proposed **Co**ntext-conditioned **B**ehavior **E**valuator (**CoBE**) .

4.1 Behavior Turing Test Formulation

BTT evaluates how closely an observed behavior matches how a human would act under the same task and scene. We formulate HuCoBench as a human-aligned contextual human-likeness score prediction benchmark conditioned on behavior, task, and scene.

For each sample, an evaluator receives a Context-SMPL behavior video $\mathcal{V}$, a textual Task Goal T, and a behavior-relevant Scene Description S, and predicts

$$\hat{y} = f_\theta(\mathcal{V}, T, S), \qquad \hat{y} \in [0, 5]. \tag{1}$$

Here, f_θ denotes the evaluator, and the reference target y is the aggregated human score described in Sec. 3.3. Higher scores indicate stronger agreement with how a human would naturally behave under the specified task and scene.

4.2 Human-Aligned Behavior Evaluators

We evaluate three model families on HuCoBench: general-purpose VLMs, human-supervised VLMs, and dedicated scoring models including CoBE. Fig. 5 illustrates the two human-aligned trainable designs: VLM adaptation with human supervision and our CoBE architecture.

General-Purpose VLMs. We first evaluate whether pretrained VLMs can perform Behavior Turing Test without task-specific training. We include proprietary models Gemini-3.5-Flash (Google DeepMind, 2026) and Qwen3.8-Max (Qwen Team, 2026e), together with InternVL3.5-8B/30B-A3B (Wang et al., 2025c) and Qwen3.5-4B/9B/27B (Qwen Team, 2026a). All models receive the same Context-SMPL video, Task Goal, and Scene Description and are prompted to output a score from 0 to 5. The evaluation protocol is provided in Appendix E.2.

Human-Supervised Fine-tuning VLMs. We next study whether direct supervision from human contextual ratings improves alignment. We fine-tune Qwen3.5-4B and Qwen3.5-9B using LoRA (Hu et al., 2022), with the pretrained backbone frozen. Each sample contains the Context-SMPL video, its Task Goal and Scene Description, and the aggregated human score as the target response. The adaptation pipeline is shown in Fig. 5(a), with training details in Appendix E.3.

Dedicated Scoring Models. For a broader comparison across evaluation paradigms, we also include ViT (Dosovitskiy et al., 2021), MotionBERT (Zhu et al., 2023), QGVL (Xu et al., 2025), and PTR-Net (Li et al., 2026), covering visual, motion-based, and visual-language scoring methods.

Context-conditioned Behavior Evaluator (CoBE). We introduce **CoBE**, a dedicated evaluator for BTT. Its core idea is that the Task Goal and Scene Description specify what behavioral evidence

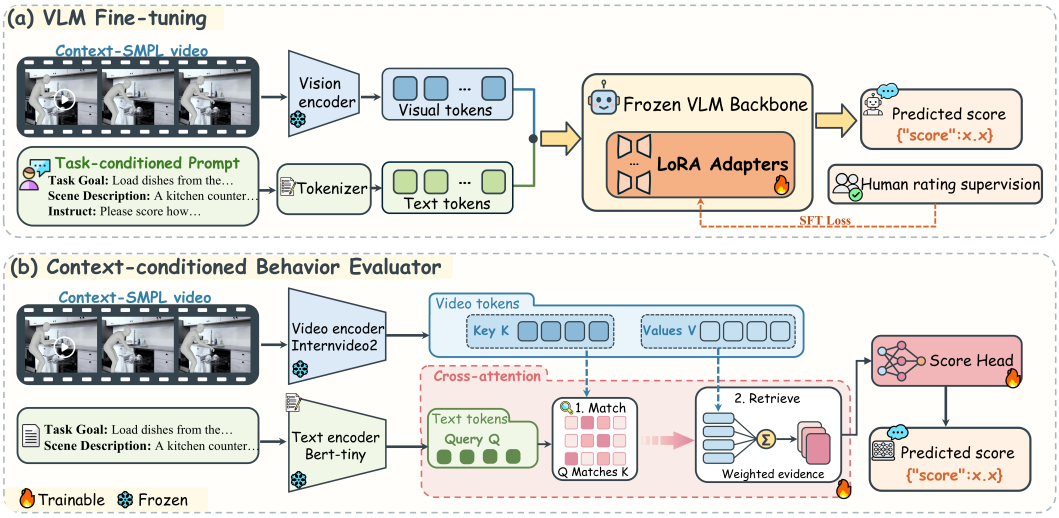

Figure 5: Human-supervised VLM adaptation and CoBE pipeline. (a) VLM fine-tuning with LoRA using contextual human ratings. (b) CoBE uses Task Goal and Scene Description as queries to retrieve relevant behavioral evidence from video tokens for score prediction.

should be examined in the video. CoBE therefore uses semantic context to retrieve task and scene-relevant evidence from video features. As shown in Fig. 5(b), it consists of a video encoder, a text encoder, a context-conditioned cross-attention module, and a scoring head.

(i) Video and Text Encoding. A frozen InternVideo2 encoder (Wang et al., 2024) extracts spatio-temporal video tokens $X_V = E_V(\mathcal{V}) \in \mathbb{R}^{N_V \times D_V}$. The Task Goal and Scene Description are concatenated and encoded by a frozen BERT-tiny encoder (Turc et al., 2020), $X_T = E_T(T, S) \in \mathbb{R}^{L \times D_T}$. Learnable projections map both token sequences into a shared attention space.

(ii) Context-Conditioned Evidence Retrieval. CoBE uses semantic context to retrieve behavioral evidence relevant to the current task and scene. The text tokens serve as queries, while the video tokens provide keys and values, with $Q = X_T W_Q$, $K = X_V W_K$, and $V = X_V W_V$. The context-conditioned attention weights are computed as

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \qquad A \in \mathbb{R}^{L \times N_V}. \tag{2}$$

and the retrieved video evidence is $F_{\text{attn}} = AV \in \mathbb{R}^{L \times d}$. This design allows the Task Goal and Scene Description to attend to visual and temporal evidence relevant to contextual judgment. The attended features are fused with the text representation via residual connection and normalization, then pooled and passed to an MLP scoring head to predict $\hat{y}$.

(iii) Human-Alignment Objective. CoBE is trained to regress the aggregated contextual human-likeness score using mean squared error:

$$\mathcal{L}_{\text{align}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2, \tag{3}$$

where y_i and $\hat{y}_i$ denote the human reference score and model prediction for the i-th sample, respectively. More architecture and training details are provided in Appendix E.4.

5 Experiments

5.1 Experimental Setup

We use the same data split for all supervised models and evaluate all methods on a shared test set. We report Spearman's rank correlation (ρ), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) against the aggregated human reference scores. Dataset splits and metric definitions are provided in Appendix E.1, with model-specific settings in Appendix E.2-E.4.

Table 2: Main results of Behavior Turing Test on HuCoBench. * denotes without Task Goal or Scene Description inputs. Best results are shown in **bold** and second-best results are underlined.

Evaluator	MAE ↓	RMSE ↓	Spearman's ρ ↑
Proprietary VLMs			
Gemini-3.5-Flash (Google DeepMind, 2026)	0.574	0.740	0.311
Qwen3.8-Max (Qwen Team, 2026e)	0.771	1.018	0.103
Open-source VLMs			
InternVL3.5-8B (Wang et al., 2025c)	0.554	0.704	0.136
InternVL3.5-30B-A3B (Wang et al., 2025c)	0.625	0.900	0.147
Qwen3.5-4B (Qwen Team, 2026c)	1.679	2.082	−0.036
Qwen3.5-9B (Qwen Team, 2026d)	0.681	0.974	0.083
Qwen3.5-27B (Qwen Team, 2026b)	1.497	1.778	−0.102
Fine-tuned open-source VLMs			
Qwen3.5-4B + LoRA (Ours) (Qwen Team, 2026c)	0.572	0.735	0.436
Qwen3.5-9B + LoRA (Ours) (Qwen Team, 2026d)	<u>0.544</u>	<u>0.698</u>	0.478
Dedicated scoring models			
ViT* (Dosovitskiy et al., 2021)	0.639	0.873	0.294
MotionBERT (frozen)* (Zhu et al., 2023)	0.656	0.856	0.220
MotionBERT (fine-tuned)* (Zhu et al., 2023)	0.620	0.803	0.312
QGVL (Xu et al., 2025)	0.551	0.741	<u>0.512</u>
PTR-Net* (Li et al., 2026)	0.635	0.827	0.243
CoBE (Ours)	**0.459**	**0.605**	**0.703**

Table 3: Counterfactual context evaluation on HuCoBench under Task-CF and Scene-CF interventions. DC denotes Direction Consistency between human and model score changes.

Evaluator	Task-CF				Scene-CF			
	ΔMAE ↓	ΔRMSE ↓	ρ_Δ ↑	DC ↑	ΔMAE ↓	ΔRMSE ↓	ρ_Δ ↑	DC ↑
Gemini-3.5-Flash	0.781	0.963	0.274	68.3%	0.642	0.811	0.231	64.5%
Qwen3.5-9B + LoRA	0.538	0.692	0.448	80.4%	0.512	0.661	0.397	76.7%
QGVL	<u>0.501</u>	<u>0.648</u>	<u>0.472</u>	<u>82.5%</u>	<u>0.468</u>	<u>0.612</u>	<u>0.431</u>	<u>78.4%</u>
CoBE (Ours)	**0.372**	**0.491**	**0.621**	**90.5%**	**0.401**	**0.526**	**0.548**	**84.2%**

Table 4: Ablation study of CoBE on HuCoBench.

Variant	MAE ↓	RMSE ↓	ρ ↑
w/o Text Branch	0.518	0.709	0.634
w/o Scene Description	<u>0.486</u>	<u>0.657</u>	<u>0.661</u>
w/o Task Goal	0.501	0.681	0.646
w/o Cross-attention	0.502	0.740	0.628
CoBE	**0.459**	**0.605**	**0.703**

5.2 Main Benchmark Results

Table 2 reports the main results on HuCoBench. General-purpose VLMs show limited agreement with human contextual judgments, and larger model variants do not consistently improve performance. Human supervision substantially improves alignment: for Qwen3.5-4B, fine-tuning increases Spearman's ρ from -0.036 to 0.436 and reduces MAE, with similar gains for Qwen3.5-9B. Among dedicated scoring models, MotionBERT (Zhu et al., 2023) and PTR-Net (Li et al., 2026) remain limited without task and scene context. QGVL (Xu et al., 2025) incorporates visual-language information and reaches $\rho = 0.512$. CoBE achieves the strongest alignment, with MAE $= 0.459$, RMSE $= 0.605$, and $\rho = 0.703$, improving ρ from 0.512 over QGVL. These results support both human contextual supervision and context-conditioned evidence retrieval for BTT.

5.3 Counterfactual Context Analysis

We use controlled interventions to isolate the effects of task and scene on BTT judgments. In **Task-Counterfactual** (Task-CF), the Context-SMPL video and Scene Description remain fixed while the Task Goal is replaced with an alternative objective. In **Scene-Counterfactual** (Scene-CF), the same SMPL-X motion and Task Goal are preserved while the motion is re-rendered in a different scene. Human scores decrease in 95.5% of Task-CF pairs and 77.5% of Scene-CF pairs (Fig. A9), showing that the same underlying motion can receive different human-likeness judgments under different task or scene conditions. Subset construction and distributions are detailed in Appendix D.2.

We then evaluate whether evaluators follow these changes. For each original-counterfactual pair, we define $\Delta y_i = y_i^{\text{orig}} - y_i^{\text{cf}}$ and $\Delta \hat{y}_i = \hat{y}_i^{\text{orig}} - \hat{y}_i^{\text{cf}}$, where positive values indicate a score decrease after intervention. Table 3 reports ρ_Δ between human and predicted changes, together with ΔMAE,

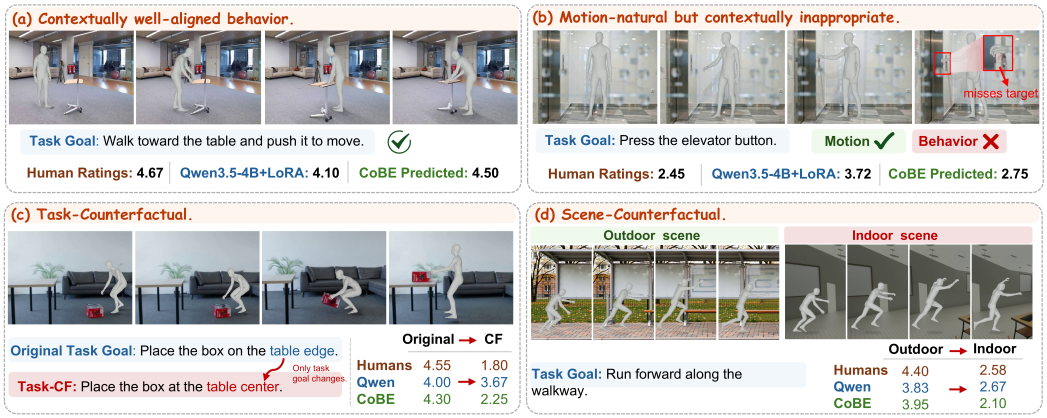

Figure 6: Qualitative analysis on HuCoBench. (a) A contextually well-aligned behavior. (b) A motion-level human-like but contextually inappropriate behavior. (c) A fine-grained Task-CF example, where a subtle Task Goal change alters the expected behavior. (d) A Scene-CF example, where the same underlying motion is evaluated in a different environment.

ΔRMSE, and **Direction Consistency (DC)**, the proportion of pairs with matching change directions. CoBE performs best across all metrics on both Task-CF and Scene-CF, showing stronger agreement with human judgments to contextual changes. Representative examples are shown in Fig. 6(c-d).

5.4 Qualitative Analysis

Fig. 6 presents representative BTT examples. In **(a)**, the behavior is well aligned with the task and scene, and CoBE closely follows the human judgment. In **(b)**, the reaching motion appears natural, yet the hand misses the elevator button. This illustrates a key distinction: human-like motion can still yield contextually inappropriate behavior. The fine-tuned VLM overestimates the score, whereas CoBE better captures this mismatch. In **counterfactual examples (c)**, a subtle Task Goal change alters the expected outcome, lowering the human score when the execution no longer matches the target. In **(d)**, the same running motion is appropriate at the outdoor bus station but less so in a closed room. CoBE follows both fine-grained task changes and broader scene changes more closely than the fine-tuned VLM. Such sensitivity is important for VLA systems, where small changes in goals or environments may require different actions.

5.5 Ablation Studies

Table 4 examines the contributions of semantic context and context-conditioned evidence retrieval. Removing either the Task Goal or Scene Description consistently degrades performance, confirming that both contribute to contextual evaluation. Removing the full text branch or replacing cross-attention with simple feature fusion further weakens alignment with human judgments. Full CoBE achieves the best performance across all three metrics.

6 Conclusion

We introduce the **Behavior Turing Test** to evaluate whether an observed behavior matches how a human would act under the same task and scene. We construct **HuCo** with Context-SMPL representations, task-scene annotations, and human contextual judgments, and establish **HuCoBench** for systematic evaluation. Our analysis reveals a key distinction between motion and behavior-level human-likeness: some humanoid behaviors can pass MTT yet fail BTT once task and scene context are considered. We further observe a persistent human-humanoid gap under contextual evaluation, highlighting the remaining challenge of context-appropriate humanoid behavior. We introduce **CoBE**, which retrieves context-relevant behavioral evidence and achieves the strongest alignment with human judgments. Overall, our framework makes the gap between *moving like a human* and *behaving like one* measurable, providing an important evaluation foundation for advancing humanoids toward context-appropriate behavior in complex real-world environments.

AI USE STATEMENT

AI tools were used for language refinement and grammar checking. Gemini 3.5 Flash was also used to assist in generating initial semantic annotations for HuCo, as described in the paper. All AI-assisted outputs used in the manuscript were reviewed and verified by the authors.

ETHICS STATEMENT

In this work, HuCo is constructed from publicly available academic datasets and publicly released humanoid demonstrations. We follow the licenses, terms of use, and redistribution policies of the original sources.

1. **Copyright and Licensing.** Copyright and ownership of the source materials remain with their respective rights holders. Any released content will follow the applicable licenses and redistribution permissions of the original sources.
2. **Privacy Protection.** We take measures to reduce privacy and identity exposure. Context-SMPL replaces the original acting subject with a standardized body representation. Before release, retained media will be reviewed for identifiable personal information, and content with unresolved privacy concerns will not be redistributed.
3. **Responsible Data Release.** HuCo is intended for non-commercial research use. Human annotation results will be released only in aggregated form, and all dataset releases will follow the corresponding source permissions and ethical requirements.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of dataset construction, annotation protocols, evaluation settings, and model training in the main paper and appendix. We will release HuCo, HuCoBench, evaluation code, and model implementations to support reproducibility and future research. Released data will follow the licenses and permissions of the original sources.

REFERENCES

Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. In *CoRL*, 2025.

Arash Amini, Hafez Farazi, and Sven Behnke. Real-time pose estimation from images for multiple humanoid robots. In *RoboCup*, 2021.

Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. CIRCLE: Capture in rich contextual environments. In *CVPR*, 2023.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *CVPR*, 2022.

Boston Dynamics. Atlas humanoid robot. `https://bostondynamics.com/products/atlas/`, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. BotChat: Evaluating LLMs' capabilities of having multi-turn dialogues. In *Findings of NAACL*, 2024.

Figure. Figure 03. `https://www.figure.ai/figure`, 2025.

Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. HumanPlus: Humanoid shadowing and imitation from humans. In *CoRL*, 2024.

Google DeepMind. Gemini 3.5 Flash. `https://deepmind.google/models/model-cards/gemini-3-5-flash/`, 2026.

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *ACM MM*, 2020.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *CVPR*, 2022.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, 2019.

Ashutosh Hathidara, Julien Yu, Vaishali Senthil, Sebastian Schreiber, and Anil Babu Ankisettipalli. Mirrorbench: A benchmark to evaluate conversational user-proxy agents for human-likeness. In *KDD Datasets and Benchmarks Track*, 2026.

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris M. Kitani, Changliu Liu, and Guanya Shi. OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *CoRL*, 2024.

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

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *CVPR*, 2024a.

Zhenyu Jiang, Yuqi Xie, Jinhan Li, Ye Yuan, Yifeng Zhu, and Yuke Zhu. Harmon: Whole-body motion generation of humanoid robots from language descriptions. In *CoRL*, 2024b.

Taehoon Kim, Jongwook Choi, Haeun Noh, Junyeup Hwang, and Jongwon Choi. I-robot: Identifying robotic and human motion in humanoids. In *ICML AI4GOOD Workshop*, 2026.

Kyungmin Lee, Sibeen Kim, Youngdo Lee, Minho Park, Hyunseung Kim, Dongyoon Hwang, Donghu Kim, Hojoon Lee, and Jaegul Choo. Phuma: Physically reliable humanoid locomotion dataset. *arXiv preprint arXiv:2510.26236*, 2025.

Yuxuan Lei, Jianxun Lian, Defu Lian, Jincenzi Wu, Tianfu Wang, and Xing Xie. A computational framework for evaluating human-likeness in llms' open-ended human behaviors. In *ICML*, 2026.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, Mona Anvari, Minjune Hwang, Manasi Sharma, Arman Aydin, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Silvio Savarese, Hyowon Gweon, Karen Liu, Jiajun Wu, and Li Fei-Fei. BEHAVIOR-1K: A benchmark for embodied AI with 1,000 everyday activities and realistic simulation. In *CoRL*, 2022.

Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics*, 2023.

Mingzhe Li, Mengyin Liu, Zekai Wu, Xincheng Lin, Junsheng Zhang, Ming Yan, Zengye Xie, Changwang Zhang, Chenglu Wen, Lan Xu, Siqi Shen, and Cheng Wang. Towards motion turing test: Evaluating human-likeness in humanoid robots. In *CVPR*, 2026.

Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 1932.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-X: A large-scale 3d expressive whole-body human motion dataset. In *NeurIPS*, 2023.

Jing Lin, Ruisi Wang, Junzhe Lu, Ziqi Huang, Guorui Song, Ailing Zeng, Xian Liu, Chen Wei, Wanqi Yin, Qingping Sun, Zhongang Cai, Lei Yang, and Ziwei Liu. The quest for generalizable motion generation: Data, model, and evaluation. In *ICLR*, 2026.

Yun Liu, Bowen Yang, Licheng Zhong, He Wang, and Li Yi. Mimicking-bench: A benchmark for generalizable humanoid-scene interaction learning via human mimicking. *arXiv preprint arXiv:2412.17730*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. HUMOTO: A 4d dataset of mocap human object interactions. In *ICCV*, 2025.

lvhaidong. LAFAN1 Retargeting Dataset. `https://huggingface.co/datasets/lvhaidong/LAFAN1_Retargeting_Dataset`, 2025.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019.

Jiageng Mao, Siheng Zhao, Siqi Song, Tianheng Shi, Junjie Ye, Mingtong Zhang, Haoran Geng, Jitendra Malik, Vitor Guizilini, and Yue Wang. Learning from massive human videos for universal humanoid pose control. In *Humanoids*, 2025.

Andre Meixner, Mischa Carl, Franziska Krebs, Noemie Jaquier, and Tamim Asfour. Towards unifying human likeness: Evaluating metrics for human-like motion retargeting on bimanual manipulation tasks. In *ICRA*, 2024.

NVIDIA. Physicalai worldmodel synthetic digital human scenes. `https://huggingface.co/datasets/nvidia/PhysicalAI-WorldModel-Synthetic-Digital-Human-Scenes`, 2026a.

NVIDIA. Physicalai worldmodel synthetic embodied robot scenes. `https://huggingface.co/datasets/nvidia/PhysicalAI-WorldModel-Synthetic-Embodied-Robot-Scenes`, 2026b.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019.

Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big Data*, 2016.

Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *CVPR*, 2021.

Qwen Team. Qwen3.5. `https://qwen.ai/blog?id=qwen3.5`, 2026a.

Qwen Team. Qwen3.5-27B model card. `https://huggingface.co/Qwen/Qwen3.5-27B`, 2026b.

Qwen Team. Qwen3.5-4B model card. `https://huggingface.co/Qwen/Qwen3.5-4B`, 2026c.

Qwen Team. Qwen3.5-9B model card. `https://huggingface.co/Qwen/Qwen3.5-9B`, 2026d.

Qwen Team. Qwen3.8-Max. `https://qwen.ai/blog?id=qwen3.8`, 2026e.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. In *ICLR*, 2025.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

ENGINEAI Robotics. ENGINEAI PM01. https://www.engineai.com.cn/product-pm01.html, 2024a.

Unitree Robotics. Unitree G1. https://www.unitree.com/cn/g1, 2024b.

Carmelo Sferrazza, Dun-Ming Huang, Xingyu Lin, Youngwoon Lee, and Pieter Abbeel. Humanoid-Bench: Simulated humanoid benchmark for whole-body locomotion and manipulation. In *RSS*, 2024.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia*, 2024.

P. E. Shrout and J. L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 1979.

Tesla. Tesla Optimus. https://www.tesla.com/AI, 2022.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2023.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-Read Students Learn Better: On the Importance of Pre-Training Compact Models. In *ICLR*, 2020.

Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOE: Real-time seeing anything. In *ICCV*, 2025a.

Haoru Wang, Wentao Zhu, Luyi Miao, Yishu Xu, Feng Gao, Qi Tian, and Yizhou Wang. Aligning human motion generation with human perceptions. In *ICLR*, 2025b.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025c.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, 2024.

Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *NeurIPS*, 2022.

Tianyi Xie, Haotian Zhang, Jinhyung Park, Zi Wang, Bowen Wen, Jiefeng Li, Xueting Li, Qingwei Ben, Haoyang Weng, Yufei Ye, David Minor, Tingwu Wang, Chenfanfu Jiang, Sanja Fidler, Jan Kautz, Linxi Fan, Yuke Zhu, Zhengyi Luo, Umar Iqbal, and Ye Yuan. GRAIL: Generating humanoid loco-manipulation from 3d assets and video priors. In *CoRL*, 2026.

Huangbiao Xu, Huanqi Wu, Xiao Ke, Yuezhou Li, Rui Xu, and Wenzhong Guo. Quality-guided vision-language learning for long-term action quality assessment. *IEEE Transactions on Multimedia*, 2025.

Ming Yan, Xin Wang, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, and Cheng Wang. CIMI4D: A large multimodal climbing motion dataset under human-scene interactions. In *CVPR*, 2023.

Ming Yan, Yan Zhang, Shuqiang Cai, Shuqi Fan, Xincheng Lin, Yudi Dai, Siqi Shen, Chenglu Wen, Lan Xu, Yuexin Ma, and Cheng Wang. RELI11D: A comprehensive multimodal human motion dataset and method. In *CVPR*, 2024.

Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C. Karen Liu, Rocky Duan, and Guanya Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. In *ICRA*, 2026.

Chenhao Yu, Hongwu Wang, Weitao Zhang, Youhao Hu, Jiachen Zhang, Gangyang Li, Alois Knoll, and Shaqi Luo. Thorarena: Benchmarking humanoid physical interaction with human motion-force demonstrations. *arXiv preprint arXiv:2607.06052*, 2026.

Kun Zhao. Retargeted amass for robotics. `https://huggingface.co/datasets/fleaven/Retargeted_AMASS_for_robotics`, 2025.

Zhenyu Zhao, Hongyi Jing, Xiawei Liu, Jiageng Mao, Abha Jha, Hanwen Yang, Rong Xue, Sergey Zakharov, Vitor Campagnolo Guizilini, and Yue Wang. Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation. In *ICRA*, 2026.

Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. ProPainter: Improving propagation and transformer for video inpainting. In *ICCV*, 2023.

Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. MotionBERT: A unified perspective on learning human motion representations. In *ICCV*, 2023.

Appendix

Appendix Overview. Thanks for reading the appendix. In this appendix file, we provide additional details in the Behavior Turing Test (BTT). Appendix A details the construction and composition of HuCo. Appendix B describes the Context-SMPL processing pipeline and quality control. Appendix C gives the semantic and human annotation protocols. Appendix D provides additional analyses of contextual and motion human-likeness, including the counterfactual subsets. Appendix E provides implementation details for HuCoBench and its evaluators. Appendix F summarizes limitations, future directions, and data usage and privacy considerations.

A HuCo Dataset Construction Details

A.1 Complete Dataset Curation Pipeline

HuCo is constructed from heterogeneous human and humanoid behavior sources through the unified data curation pipeline shown in Fig. A1. We collect more than 30 hours of raw videos spanning four domains: real humans, digital humans, real humanoid robots, and simulated humanoids. We temporally segment the videos and retain clips with a clear task objective and meaningful interaction with the surrounding environment. Each clip contains a coherent task execution sequence lasting 5-30 seconds. We exclude context-free motions, incomplete behaviors, and clips without sufficient task or scene information. After filtering, 4.3 hours of behavior clips are retained for subsequent processing.

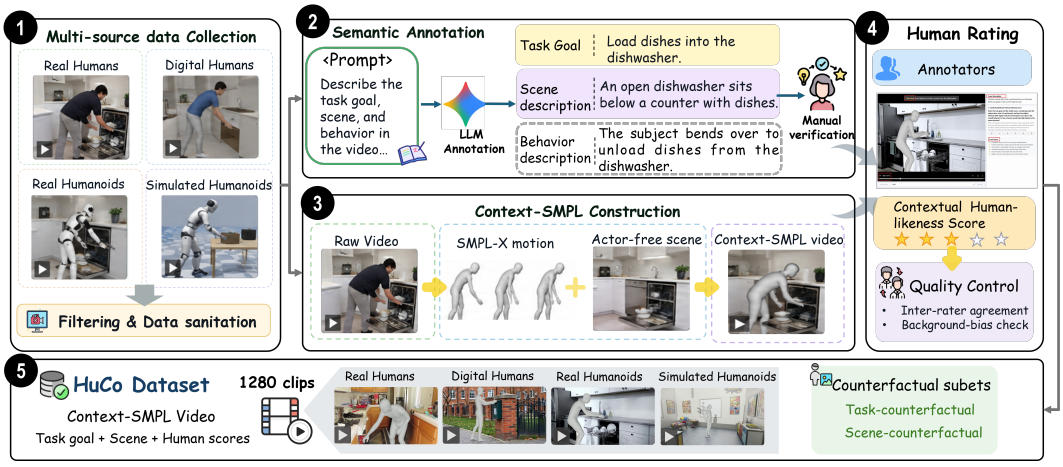

Figure A1: HuCo data curation pipeline. Each selected clip undergoes semantic annotation and Context-SMPL construction in parallel, followed by contextual human-likeness scoring and quality control.

Each selected clip is processed along two parallel paths. For semantic annotation, the original video is used as the primary evidence to generate the *Task Goal*, *Scene Description*, and an auxiliary *Behavior Description*, while source labels serve only as references. The *Behavior Description* summarizes the observed actions for manual verification and is not provided to human raters or BTT evaluators. The annotator is instructed to ignore embodiment-specific appearance and focus on the observed behavior and its context, as detailed in Appendix C.1. In parallel, the same clip is converted into the unified Context-SMPL representation described in Sec. 3.2 and Appendix B. This representation suppresses embodiment-specific appearance cues while preserving the motion and scene context required for behavior evaluation. The reconstructed clips then undergo stage-wise quality control before human annotation.

Human annotators then evaluate each Context-SMPL clip using its *Task Goal* and *Scene Description*. Valid ratings are averaged to obtain the contextual human-likeness score used as the human reference in HuCoBench. After semantic verification, Context-SMPL quality control, and annotation quality checks, the final HuCo dataset contains 1,280 behavior clips totaling 4.3 hours.

A.2 Data Sources and Composition

HuCo contains 1,280 behavior clips evenly distributed across four domains: Real Human, Digital Human, Real Humanoid, and Simulated Humanoid, with 320 clips per domain (Table A1). This yields 640 human and 640 humanoid clips in total. Real Human samples are drawn from HUMOTO (Lu et al., 2025), BEHAVE (Bhatnagar et al., 2022), and the in-the-wild subset of ViMoGen-228K (Lin et al., 2026). We retain clips with clear task intent, coherent execution, and interpretable interaction with the surrounding scene. Digital Human samples are collected from NVIDIA's Synthetic Digital Human Scenes (NVIDIA, 2026a) and the generated-motion subset of ViMoGen-228K (Lin et al., 2026), providing additional diversity in behaviors and environments.

Table A1: Composition and curation statistics of HuCo. Raw Data denotes the collected duration before filtering, while Retained Data denotes the duration of the final curated clips.

Group	Domain	Raw Data (h)	Retained Data (h)	#Clips	Share (%)
Human	Real Human	7.6	1.12	320	25
Human	Digital Human	6.8	1.07	320	25
Humanoid	Real Humanoid	9.8	1.03	320	25
Humanoid	Simulated Humanoid	6.7	1.08	320	25
Total	–	30.9	4.30	1,280	100

The humanoid group contains both real-robot demonstrations and simulated behaviors. For the Real Humanoid domain, we collect publicly released demonstrations covering 21 humanoid robot models. Representative systems include Figure 03 (Figure, 2025), Tesla Optimus (Tesla, 2022), Boston Dynamics Atlas (Boston Dynamics, 2024), Unitree G1 (Robotics, 2024b), and ENGINEAI PM01 (Robotics, 2024a). The collected videos cover goal-directed navigation, object handling, tool and environment interaction, and other task-oriented whole-body behaviors. For the Simulated Humanoid domain, samples are drawn from GRAIL (Xie et al., 2026) and NVIDIA's Synthetic Embodied Robot Scenes (NVIDIA, 2026b). We visualize the corresponding motion sequences in their associated environments and temporally crop them into coherent behavior clips following the same curation criteria used for the other domains. Fig. A2 shows original and Context-SMPL examples from five representative scene categories.

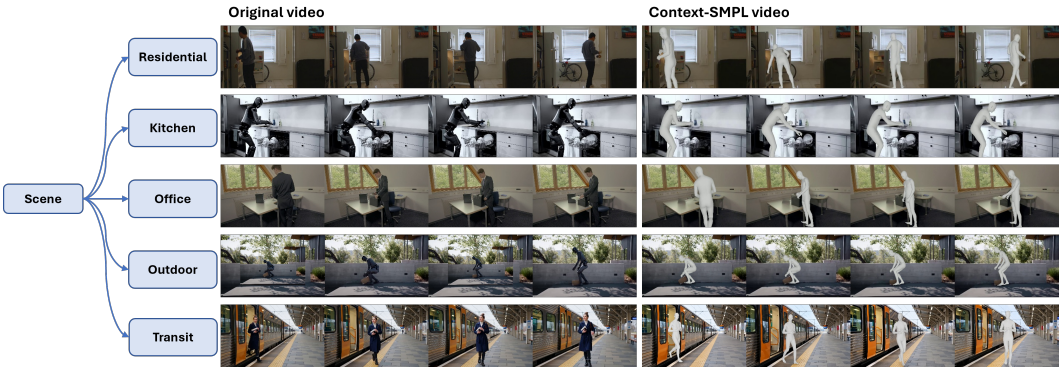

Figure A2: Representative HuCo samples from five scene categories. Each row shows original video frames and the corresponding Context-SMPL representation.

A.3 Data Filtering and Clip Standardization

Clip Extraction and Standardization. For long-form source videos, we extract temporally localized segments that contain a coherent task execution sequence. Each clip preserves the behavioral progression needed to infer the task objective and its interaction with the surrounding environment. The retained clips last 5-30 seconds and are standardized to 30 fps before semantic annotation and Context-SMPL reconstruction. This provides a consistent temporal representation across heterogeneous source datasets.

Processing Quality Screening. We apply stage-wise screening during Context-SMPL processing to ensure that each retained clip can be represented reliably. A sample is kept only when the acting subject is correctly localized, the recovered motion preserves the major behavior, and the task-relevant scene context remains interpretable. Screening is performed before human scoring and is based solely on representation fidelity. Detailed quality-control criteria are provided in Appendix B.4.

Manual Verification. Before inclusion in HuCo, each processed sample undergoes a final manual review. We verify that the clip contains a coherent task-directed behavior, that its Task Goal and Scene Description are consistent with the visible evidence, and that the Context-SMPL video preserves the behavioral and scene information required for contextual judgment. Samples with unresolved semantic ambiguity or substantial processing errors are excluded.

B Context-SMPL Construction Details

B.1 Subject Segmentation and Background Reconstruction

For an input sequence $\{I_t\}_{t=1}^{F}$, we first estimate an actor mask M_t for each frame. We use Grounded-SAM with SAM 2 (Ren et al., 2024; Ravi et al., 2025) for subject segmentation and ProPainter (Zhou et al., 2023) for video inpainting, producing an actor-free background sequence $\{B_t\}_{t=1}^{F}$. We manually inspect the segmentation and inpainting results to ensure that the acting subject is correctly removed and that behavior-relevant scene content is preserved.

We further refine subject localization using YOLOE (Wang et al., 2025a) detections together with SAM 2 (Ravi et al., 2025) masks. Detection results are manually corrected when necessary, and the corresponding subject masks are regenerated before video inpainting. We then perform a second quality check on the reconstructed backgrounds and remove samples with persistent segmentation errors or severe artifacts in task-relevant regions.

B.2 SMPL-X Motion Recovery

In parallel with background reconstruction, we recover the acting subject as an SMPL-X motion sequence (Pavlakos et al., 2019) using the motion-recovery pipeline based on GVHMR (Shen et al., 2024). The recovered representation contains frame-wise body pose together with global translation and orientation, providing a common parametric motion representation across human and humanoid samples. We retain both local body articulation and global trajectory, since contextual behavior evaluation depends on how the subject moves relative to surrounding objects and the scene.

For samples with unreliable recovery, we inspect subject localization and tracking before reprocessing. In humanoid clips where person-based localization fails, we use YOLOE (Wang et al., 2025a) with the target category set to `robot` to obtain corrected subject tracks, then regenerate the corresponding masks and rerun motion recovery. Reconstructed sequences are manually checked for major pose errors, trajectory drift, and incorrect interaction geometry. Samples are retained only when the recovered motion preserves the action trajectory and spatial interaction pattern needed to interpret the behavior.



Figure A3: Intermediate stages of Context-SMPL construction from the original video to the final re-rendered representation.

B.3 Context-SMPL Re-rendering

The recovered SMPL-X motion is re-rendered into the corresponding actor-free background to construct the final Context-SMPL video, as shown in Fig. A3. We preserve the original temporal ordering and align the reconstructed body with the scene so that its global trajectory, scale, and scene-relative position remain consistent with the observed behavior. This preserves spatial relationships that are important for contextual judgment. The resulting Context-SMPL representation retains the body motion, its spatial relation to the environment, and the scene context needed for BTT, while presenting human and humanoid behaviors in a unified visual form.

B.4 Quality Assurance and Exclusion Criteria

Context-SMPL construction involves segmentation, inpainting, motion recovery, and re-rendering. We therefore apply stage-wise quality control followed by a final manual review. A sample is retained only when the processed video preserves the task-relevant behavioral and scene information of the original clip with sufficient fidelity for contextual evaluation.

We inspect four major sources of processing error: (1) incomplete subject removal or visible actor residuals; (2) background reconstruction artifacts that corrupt task-relevant objects or interaction regions; (3) substantial SMPL-X pose or trajectory errors that alter the observed behavior; and (4) incorrect alignment between the rendered body and the scene, including severe penetration, floating, or inconsistent object-relative positioning. Samples containing these errors at a level that affects behavior interpretation are excluded from HuCo. Quality control is performed before human scoring and is based on the fidelity of the Context-SMPL reconstruction.

C Annotation Protocol and Quality Control

C.1 Semantic Annotation Prompt

Each HuCo clip is annotated with a *Task Goal* and a *Scene Description*, which provide the semantic context for BTT. The *Task Goal* describes the intended objective, while the *Scene Description* summarizes environmental information relevant to the observed behavior. We use Gemini 3.5 Flash (Google DeepMind, 2026) to generate initial annotations from the original video, sampled at 5 fps, together with the available source action category, behavior description, and scene label. The complete annotation prompt is shown in Fig. A4.

The video serves as the primary evidence, while source labels are used only as references. The prompt instructs the model to ignore the subject's identity, embodiment, and appearance, refer to the actor uniformly as "the subject," and focus on the intended task, observed actions, and behavior-relevant scene context. We additionally generate auxiliary semantic metadata, including a *Behavior Description*, for manual verification and dataset analysis. These auxiliary fields are not shown to human raters and are not used as inputs to BTT evaluators. All generated annotations are manually reviewed, and ambiguous or inaccurate entries are corrected before inclusion in HuCo.

Textual annotation prompt.

You are an expert in human, humanoid, and embodied behavior understanding.

You will receive:
One behavior video.
The original action category.
The original behavior description.
The original scene label.
The original labels are provided only as references and may be incomplete, overly general, inconsistent, or inaccurate.
Evidence Rules
Use the visible video as the primary evidence for the observed behavior and scene description.
Use the visible video as the primary evidence and the source labels as supporting references when inferring the intended task goal.
If the source labels clearly conflict with the visible evidence, follow the visible evidence.
Always refer to the acting subject as "the subject".
Do not identify or mention the subject's embodiment type, identity, appearance, or specific model.
Do not evaluate behavior quality, naturalness, motion smoothness, task performance, or human-likeness.
Generate exactly three fields:
1. task_goal
Write one short English imperative sentence describing what the subject is intended to accomplish.
Describe only the task objective.
Do not describe motion quality, success, failure, or the complete execution process.
2. behavior_description
Write one or two objective English sentences describing what is visibly performed from the beginning to the end of the video.
Describe the main action phases in chronological order.
Describe only observed behavior, including incomplete attempts when visible.
Do not infer information that is not visually supported.
Do not use evaluative terms such as natural, unnatural, smooth, awkward, clumsy, good, bad, successful, or human-like.
Always refer to the acting subject as "the subject".
3. scene_description
Write one short English sentence describing only the scene elements and spatial relationships relevant to behavior selection, navigation, or interaction.
Do not mention whether the scene is real, simulated, rendered, or generated.
Do not describe visual style, lighting, textures, video quality, subject appearance, task success, or behavior quality.
Important Distinctions
task_goal describes what the subject is intended to accomplish.
behavior_description describes what is actually visible.
scene_description describes the task-relevant environment.
The three text fields should contain complementary information and should not repeat the same sentence.
If the exact identity of an object is unclear, use a general term such as "object" or "item".
Original Labels
Action category:
{action_category}
Behavior description:
{behavior_description}
Scene:
{scene}
Return only a valid JSON object in the following format:

```
{
"task_goal": "...",
"behavior_description": "...",
"scene_description": "..."
}
```

Figure A4: Prompt for semantic annotation.

C.2 Human Annotation Interface and Scoring Rubric

We recruit 20 annotators with relevant background knowledge. For each sample, annotators are presented with the Context-SMPL video together with its *Task Goal* and *Scene Description*. The original embodiment type, source dataset, and human/humanoid category are hidden from annotators, and samples are presented in randomized order. The annotation interface is shown in Fig. A5.

Our primary question measures *Contextual Human-Likeness Score*, the target of BTT. Annotators judge how closely the observed behavior matches how a real human would naturally act to accomplish the same task in the given scene. The judgment considers action selection, task execution, object interaction, spatial relationships, and motion execution. Each sample receives an integer score from 0 to 5 according to the rubric in Table A2. Valid ratings are averaged to obtain the contextual human-likeness score used as the human reference in HuCoBench.

For comparison with motion-level evaluation, we additionally collect *Motion Human-Likeness Score*, following the Motion Turing Test protocol (Li et al., 2026), which annotators observe only the reconstructed SMPL-X (Pavlakos et al., 2019) motion without task or scene. Motion Human-Likeness score focuses on whether the body movement itself appears human-like. These ratings are used only for the contextual-versus-motion analysis in Appendix D.1; the HuCoBench prediction target is the contextual human-likeness score.

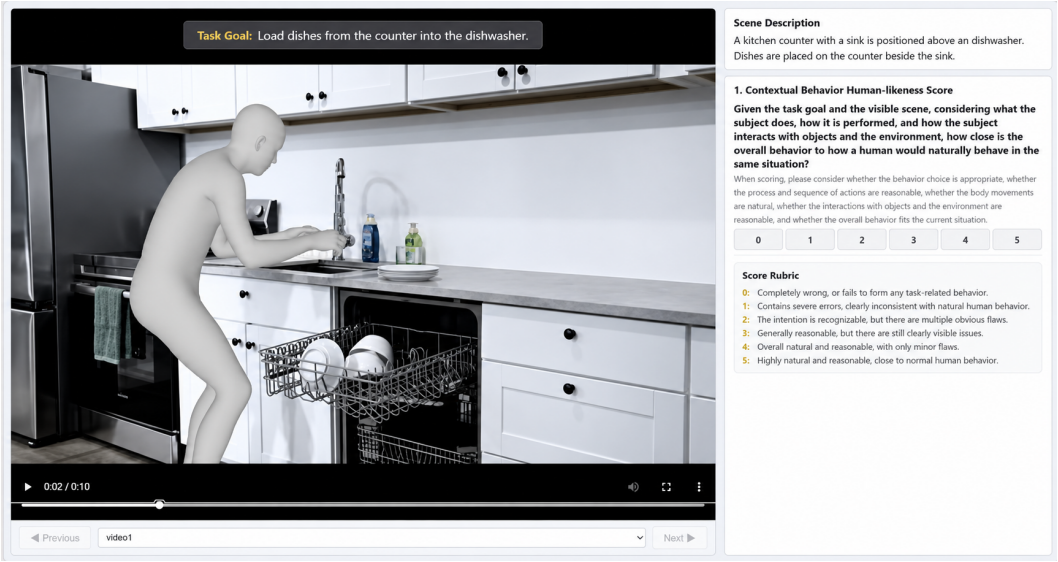

Figure A5: Human annotation interface for BTT. Annotators view the Context-SMPL video together with its Task Goal and Scene Description and assign a contextual human-likeness score.

Table A2: Scoring rubric for contextual human-likeness score in BTT.

Score	Criterion
0	Completely wrong, or fails to form any task-related behavior.
1	Contains severe errors, clearly inconsistent with natural human behavior.
2	The intention is recognizable, but there are multiple obvious flaws.
3	Generally reasonable, but there are still clearly visible issues.
4	Overall natural and reasonable, with only minor flaws.
5	Highly natural and reasonable, close to normal human behavior.

C.3 Annotation Reliability

For each behavior sample, valid human ratings are averaged to obtain the contextual human-likeness score used in HuCoBench. We assess inter-rater reliability using ICC(2,N) (Shrout & Fleiss, 1979), a two-way random-effects, absolute-agreement intraclass correlation for the mean of N ratings. This measure evaluates the reliability of the aggregated human scores while accounting for both sample-level variation and systematic differences across annotators.

Let MS_R, MS_C, and MS_E denote the sample, annotator, and residual mean squares, respectively. Let n denote the number of evaluated samples and N the number of ratings aggregated per sample. ICC(2,N) is computed as

$$\mathrm{ICC}(2, N) = \frac{MS_R - MS_E}{MS_R + (MS_C - MS_E)/n}. \tag{4}$$

As reported in Table A3, the contextual human-likeness annotations achieve an ICC(2,N) of 0.887, with a 95% confidence interval of [0.871, 0.901]. This indicates high reliability of the aggregated human judgments used as reference scores in HuCoBench.

Table A3: Inter-rater reliability of contextual human-likeness score annotations.

Measure	Estimate
ICC(2,N)	0.887
95% confidence interval	[0.871, 0.901]

Table A4: Background-only source recognition. Each row reports the predicted source distribution conditioned on the true source. Random guessing gives 50% for each prediction.

True Source	Pred. Human (%)	Pred. Humanoid (%)
Human	53.13	46.87
Humanoid	57.06	42.94
Balanced Accuracy: 48.035%	(Chance: 50.00%)	

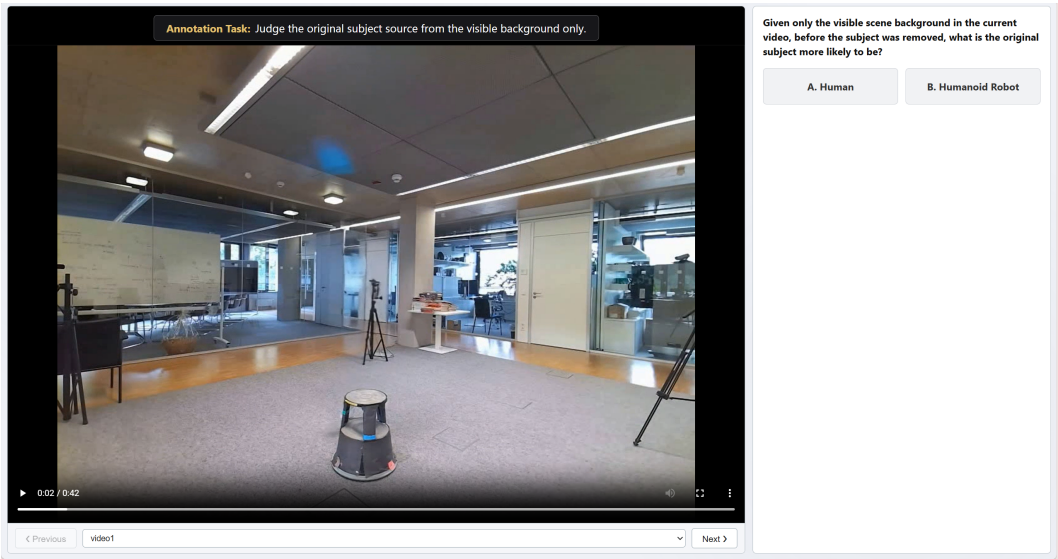

Figure A6: Background-only classification task interface.

C.4 Background-Only Bias Control

A potential concern is that the preserved background may contain domain-specific visual cues associated with human or humanoid samples. We therefore conduct a background-only control study to test whether the source domain can be inferred without observing the acting subject. We construct a subset in which the subject is removed and only the reconstructed background video is shown to annotators. Annotators perform a binary source-classification task, as illustrated in Fig. A6.

If the background introduces no systematic source bias, predictions should approach a 50%/50% distribution for both true source classes. As shown in Table A4, Human backgrounds are classified as Human and Humanoid in 53.13% and 46.87% of cases, respectively. For Humanoid backgrounds, the corresponding proportions are 57.06% and 42.94%. The resulting balanced accuracy is 48.035%, close to the 50% chance level. These results indicate that the preserved background does not provide a reliable shortcut for distinguishing Human from Humanoid samples and does not provide a reliable source-discriminative shortcut.

D Additional Dataset Analysis

D.1 Contextual vs. Motion Human-Likeness

To examine the distinction between behavior-level and motion-level human-likeness, we compare the two human judgments collected in HuCo. *Contextual Human-Likeness Score* (Q1) is the reference score used in BTT and evaluates how closely the observed behavior matches how a human would naturally act under the same Task Goal and Scene Description. *Motion Human-Likeness Score* (Q2) follows the Motion Turing Test protocol (Li et al., 2026) and evaluates whether the

Table A5: Contextual human-likeness scores (Q1) for BTT and motion human-likeness scores under the MTT protocol (Q2) across four domains. Both scores range from 0 to 5.

Domain	Q1 mean	Q2 mean	Q1–Q2	Spearman ρ
Real Human	4.347	4.457	−0.110	0.864
Digital Human	4.019	4.189	−0.170	0.927
Simulated Humanoid	3.566	3.649	−0.083	0.831
Real Humanoid	2.905	3.219	−0.314	0.957

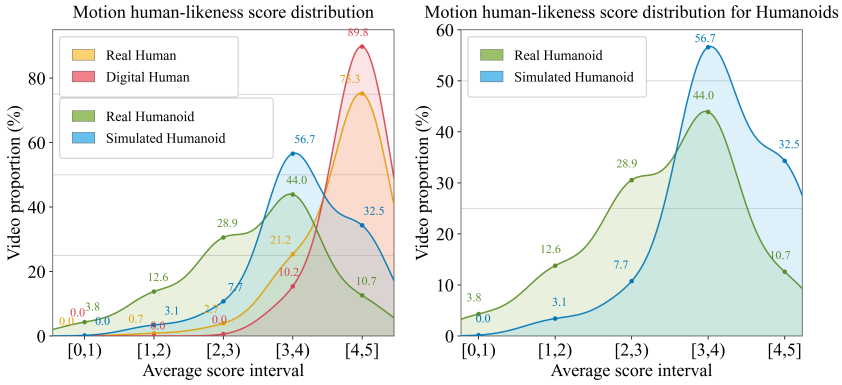

Figure A7: Distribution of motion human-likeness scores (Q2). Left: all four HuCo domains. Right: Real and Simulated Humanoids.

body movement itself appears human-like. Table A5 reports the domain-wise statistics for both judgments.

Across all four domains, Q1 and Q2 remain strongly correlated, with Spearman's ρ ranging from 0.831 to 0.957. The mean contextual score is lower than the corresponding motion score in every domain, with the largest gap appearing for Real Humanoid behaviors. Using the Real Human mean scores as human-reference criteria (4.457 for MTT and 4.347 for BTT), we further observe behaviors that exceed the MTT reference but fall below the BTT reference. A representative case is shown in Fig. 4(a). Such cases provide concrete evidence that motion judged as human-like does not necessarily correspond to human-like behavior under the intended task and scene.

Fig. A7 further shows the distribution of motion human-likeness scores across domains. Both human domains are concentrated in the highest score range, while humanoid behaviors span a broader distribution. Simulated Humanoids also receive higher motion ratings than Real Humanoids overall. Compared with the contextual score distributions in Fig. 4, the broad domain ordering remains similar, while contextual evaluation further reflects whether the behavior fits the intended task and scene.

D.2 Counterfactual Subset Construction

We construct two counterfactual subsets to isolate the effects of task and scene context on BTT judgments. In **Task-CF**, the Context-SMPL video and Scene Description are kept fixed, while the Task Goal is replaced with a counterfactual objective that changes the expected behavior or outcome. In **Scene-CF**, the same underlying SMPL-X motion and Task Goal are preserved, while the motion is re-rendered in an alternative scene and the Scene Description is updated accordingly. Task-CF therefore isolates changes in task intent, while Scene-CF isolates changes in environmental context under the same motion. Representative interventions are shown in Fig. A8.

For each original-counterfactual pair, we define the human score change as $\Delta y = y^{\text{orig}} - y^{\text{cf}}$, where positive values indicate that contextual human-likeness decreases after the intervention. As shown in Fig. A9, contextual human-likeness scores decrease for 95.5% of Task-CF pairs and 77.5% of Scene-CF pairs. These results show that changing either the intended task or the surrounding scene

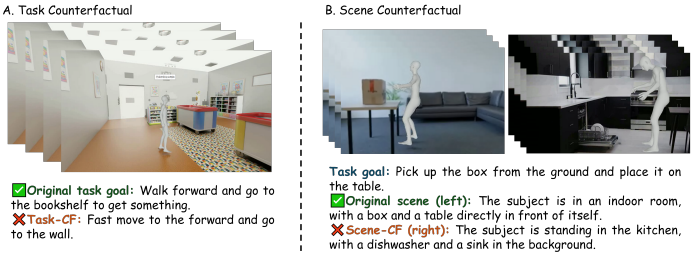

Figure A8: Representative examples of Task-CF and Scene-CF interventions.

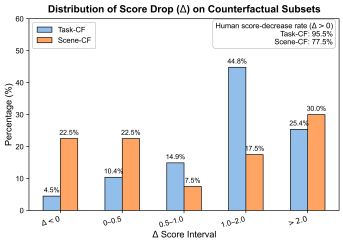

Figure A9: Distribution of human score changes under counterfactual interventions.

can alter the perceived human-likeness of the same underlying motion. Therefore, motion alone is insufficient for behavior-level human-likeness evaluation.

E HuCoBench Implementation Details

E.1 Dataset Split and Evaluation Metrics

We partition HuCo into training and test sets using a fixed 80/20 split. The same split is used across all supervised evaluators to ensure direct comparison. Aggregated contextual human-likeness scores serve as supervision during training and as reference targets for evaluation. General-purpose VLMs are evaluated directly on the same test set without HuCo-specific training.

We evaluate agreement between model predictions and human judgments using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Spearman's rank correlation coefficient (ρ). Let N denote the number of evaluated samples, y_i the aggregated contextual human-likeness score for sample i, and $\hat{y}_i$ the corresponding model prediction. All scores are evaluated on the 0-5 scale.

Mean Absolute Error (MAE). MAE measures absolute score difference between predicted and human scores:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|\,. \tag{5}$$

Lower MAE indicates closer numerical agreement with human judgments.

Root Mean Squared Error (RMSE). RMSE measures squared prediction error:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \tag{6}$$

RMSE places greater weight on large prediction errors, with lower values indicating better agreement.

Spearman's Rank Correlation (ρ). Spearman's ρ measures agreement between the rankings induced by model predictions and human scores:

$$\rho = \mathrm{corr}\left(\mathrm{rank}(y), \mathrm{rank}(\hat{y})\right), \tag{7}$$

where tied scores receive their average rank. Higher ρ indicates stronger agreement with the relative ordering of human judgments.

E.2 VLM Evaluation Protocol

We evaluate all general-purpose VLMs under a unified protocol. Each model receives the same Context-SMPL video, *Task Goal*, and *Scene Description*, together with the evaluation prompt shown in Fig. A10. The prompt defines contextual human-likeness on the 0-5 scale and asks the model to judge how closely the observed behavior matches how a human would naturally act under the same task and scene. Models are required to return a single numerical score in a fixed output format.

VLM Evaluation prompt.

You are evaluating how closely the observed behavior matches how a human would naturally act under the given Task Goal and Scene Description.

Task Goal:
{task_goal}
Scene Description:
{scene_description}

Watch the complete video and assign one continuous score from 0 to 5.
Evaluate how close the overall behavior is to how a human would naturally behave when pursuing the same task under the same scene constraints.

Consider the behavior as a complete process, including:
- Action selection: whether the chosen actions and behavioral strategy are appropriate and natural for achieving the task.
- Execution process: whether the actions are carried out in a natural, coherent, coordinated, and purposeful manner.
- Environment interaction: whether the agent responds to objects, spatial relationships, interaction targets, and scene constraints in a natural and context-appropriate way.

The score should reflect the overall behavioral human-likeness rather than task success alone. A behavior may complete the task yet receive a lower score if its action selection, execution, or interaction with the scene differs substantially from how a human would naturally behave under the same conditions.

Use the visible video as the primary evidence. Use the Task Description and Scene Description to understand the intended goal and relevant contextual constraints.

The acting subject is represented using a standardized SMPL body. Do not evaluate visual appearance, body shape, SMPL rendering quality, reconstruction quality, video quality, or camera style. Do not infer whether the original agent is a human, humanoid robot, simulated agent, or generated character. Do not use background or dataset-specific visual cues to infer the source of the behavior.

Score rubric:
0 = Completely wrong, or fails to form any task-related behavior.
1 = Contains severe errors, clearly inconsistent with natural human behavior.
2 = The intention is recognizable, but there are multiple obvious flaws.
3 = Generally reasonable, but there are still clearly visible issues.
4 = Overall natural and reasonable, with only minor flaws.
5 = Highly natural and reasonable, close to normal human behavior.
Intermediate decimal scores are allowed.

Return exactly one JSON object:
{"score": 3.8}
Do not output an explanation, reasoning, markdown, error category, recommendation, or any other key or text.

Figure A10: Prompt used for VLM evaluation on HuCoBench.

E.3 Human-Supervised VLM Fine-Tuning

We fine-tune Qwen3.5-4B (Qwen Team, 2026c) and Qwen3.5-9B (Qwen Team, 2026d) as BTT evaluators using aggregated contextual human-likeness scores as supervision. Each training sample consists of a Context-SMPL video, its *Task Goal* and *Scene Description*, and the corresponding human score in $[0, 5]$. Videos are sampled at 2 fps with at most 32 frames, and the model is trained to generate a structured response in the format $\{\texttt{"score":}\ \texttt{x.x}\}$.

We apply LoRA (Hu et al., 2022) while keeping the pretrained backbone frozen. LoRA dropout is set to 0.05. Training uses causal language-modeling loss computed only on the assistant response tokens, with a per-device batch size of 1, eight-step gradient accumulation, 5 training epochs, and an initial learning rate of 8×10^{-5}.

E.4 CoBE Architecture and Training

Architecture. CoBE is a dedicated evaluator for the Behavior Turing Test (BTT). Given a Context-SMPL video $\mathcal{V}$, a Task Goal T, and a Scene Description S, it predicts a contextual human-likeness score $\hat{y} \in [0, 5]$. The key design principle is to evaluate the observed behavior *conditioned on* the intended task and surrounding scene, rather than assessing motion quality alone.

The Context-SMPL video is first encoded by a frozen InternVideo2 encoder (Wang et al., 2024) into a sequence of spatio-temporal video tokens,

$$X_V = E_V(\mathcal{V}), \qquad X_V \in \mathbb{R}^{N_V \times D_V}. \tag{8}$$

These tokens provide the visual evidence needed for behavior evaluation, including the subject's motion, its temporal evolution, and its relation to the surrounding scene and objects. In parallel,

the Task Goal and Scene Description are concatenated and encoded by a frozen BERT-tiny encoder (Turc et al., 2020),

$$X_T = E_T(T, S), \qquad X_T \in \mathbb{R}^{L \times D_T}. \tag{9}$$

The text branch specifies *how the observed behavior should be examined*. For example, when the task is to place a box on the table to the right, the evaluator should focus on evidence such as the box trajectory, its final placement, and its spatial relation to the target table. Therefore, the textual context is used to determine which parts of the video are relevant to the current evaluation.

Context-Conditioned Evidence Retrieval. To explicitly model this conditional evaluation process, CoBE adopts a text-as-query cross-attention mechanism. Learnable projections map the two modalities into a shared attention space:

$$Q = X_T W_Q, \qquad K = X_V W_K, \qquad V = X_V W_V. \tag{10}$$

Here, the text tokens serve as queries, while the same video token sequence provides both keys and values through different learned projections. The query-key interaction first identifies video evidence relevant to the current Task Goal and Scene Description:

$$A = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \qquad A \in \mathbb{R}^{L \times N_V}, \tag{11}$$

where d is the attention dimension and the softmax is taken over video tokens for each text query.

The resulting attention weights are then used to retrieve the corresponding visual evidence from the video values:

$$F_{\mathrm{attn}} = AV, \qquad F_{\mathrm{attn}} \in \mathbb{R}^{L \times d}. \tag{12}$$

In other words, Q matches against K to determine *where* task-relevant evidence appears in the video, while V carries the visual information that is subsequently retrieved. In practice, we use multi-head cross-attention to capture multiple task-relevant relations in parallel.

Context-Aware Fusion and Scoring. The retrieved visual evidence is fused with the projected text representation through a residual connection and normalization:

$$\widetilde{X} = \mathrm{LayerNorm}\left(Q + F_{\mathrm{attn}}\right). \tag{13}$$

We then apply masked mean pooling over valid text tokens to obtain a compact behavior representation,

$$F_{\mathrm{behavior}} = \mathrm{MaskedMeanPool}(\widetilde{X}), \qquad F_{\mathrm{behavior}} \in \mathbb{R}^{d}. \tag{14}$$

Finally, a lightweight two-layer MLP maps this representation to a scalar score. A sigmoid function constrains the prediction to the same 0-5 range as the human annotations:

$$\hat{y} = 5 \cdot \sigma\left(\mathrm{MLP}(F_{\mathrm{behavior}})\right). \tag{15}$$

This design explicitly compares the semantic expectation defined by the Task Goal and Scene Description with the behavioral evidence observed in the video. Changing the task or scene changes the textual queries and can therefore shift which video evidence is emphasized, even when the underlying motion is unchanged. This mechanism directly supports the context sensitivity required by Task-CF and Scene-CF evaluation.

Training Objective. InternVideo2 and BERT-tiny remain frozen throughout training. We optimize the modality projections, multi-head cross-attention module, normalization layers, and scoring head. CoBE is trained to regress the aggregated contextual human-likeness score using mean squared error:

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2, \tag{16}$$

where $y_i \in [0, 5]$ is the aggregated human score for sample i and $\hat{y}_i$ is the corresponding prediction.

Training Configuration. We optimize all trainable components with AdamW (Loshchilov & Hutter, 2019) for 30 epochs using a batch size of 2, a constant learning rate of 10^{-4}, and weight decay of 10^{-4}.

E.5 Dedicated Scoring Baselines

We additionally evaluate ViT (Dosovitskiy et al., 2021), MotionBERT (Zhu et al., 2023), QGVL (Xu et al., 2025), and PTR-Net (Li et al., 2026). ViT operates on visual features without textual task or scene inputs. MotionBERT and PTR-Net operate on motion representations and therefore do not observe task or scene context. QGVL receives visual-language inputs and is adapted for contextual score prediction. All supervised baselines use the same training split and are evaluated on the shared test set.

F Limitations and Data Statement

F.1 Limitations and Future Work

Limitations. HuCo currently covers a finite set of robot platforms, tasks, and scene types. Extending the benchmark to broader domains and longer-horizon interactions would further strengthen evaluation under unconstrained real-world conditions. Context-SMPL relies on motion recovery and re-rendering, which can lose fine-grained contact information and morphology-dependent interaction details. Improving body-scene reconstruction and interaction fidelity is therefore an important direction for future dataset expansion.

Limitations. HuCo focuses on representative robot platforms, task types, and scene categories that support contextual behavior evaluation. Future extensions may broaden the benchmark to longer-horizon behaviors and more diverse real-world settings. Context-SMPL provides a unified representation for comparing human and humanoid behaviors while preserving task-relevant motion and scene context; further advances in reconstruction could enable even finer analysis of contact and morphology-dependent interactions.

Future Work. Future work will extend BTT to more diverse and long-horizon real-world interactions and develop stronger context-aware behavior evaluators. Beyond evaluation, CoBE can also provide human-aligned feedback for behavior generation and policy learning. We view HuCoBench as an evaluation foundation for guiding humanoid behaviors from motion-level realism toward task and scene-appropriate interaction in real-world environments.

F.2 Data Usage and Privacy Statement

HuCo is intended for non-commercial research use. Copyright of the original source materials remains with the respective rights holders, and any released content will follow the applicable licenses and permissions of its source. Context-SMPL replaces the original acting subject with a standardized body representation, reducing the exposure of appearance and identity information while preserving behavior and scene context. Before release, retained media will be reviewed for identifiable personal information, and human annotation results will be provided only in aggregated form.