AFFECTSIM: A CONTROLLABLE INTERACTIVE 3D SIMULATION BENCHMARK FOR EMBODIED AFFECTIVE PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing affective benchmarks largely rely on fixed recordings, where observation conditions are predetermined before inference. Consequently, they mainly evaluate passive affect recognition while overlooking a key question for embodied agents: how to actively acquire informative affective evidence. We introduce **AffectSim**, a controllable interactive 3D simulation benchmark for embodied affective perception. AffectSim represents affective behaviors as replayable 3D episodes and separates the underlying behavior from how it is observed, enabling systematic control over distance, orientation, occlusion, scene geometry, and agent viewpoint. It contains 27,647 episodes across five emotion categories and 57 scenes. Across 24 frozen perception-model configurations, we show that observation quality substantially affects recognition. To examine whether active sensing can recover this gap, we establish a training-free heuristic baseline, which improves recognition in 21 of 24 settings. Building on this, we propose Evidence-Aware Observation Gate (EAOG) to adaptively assess whether additional observation is still beneficial. These results show that affective perception depends not only on the observed behavior, but also on how evidence is acquired. By making observation acquisition experimentally controllable, AffectSim provides a foundation for studying embodied affective perception in interactive 3D environments. All benchmark assets, code, and simulation environments will be released.

1 INTRODUCTION

Robots operating around people must understand not only objects and instructions, but also human affect (Narayanan et al., 2020). Yet most emotion-recognition benchmarks evaluate observations that have already been collected and framed before inference (Poria et al., 2019; Wang et al., 2022). This implicitly assumes an *ideal photographer* who places the camera at an informative distance and orientation and makes the relevant behavior visible. Such an assumption does not hold for embodied agents: as illustrated in Figure 1, the same affective behavior may appear ambiguous from afar, remain partially occluded after approaching, and only become informative after active repositioning. For embodied affective perception, a good viewpoint is not given; it must be acquired.

This limitation reflects the prevailing collect-annotate-recognize paradigm in affective computing. Once affective behavior is recorded, viewpoint, distance, orientation, occlusion, and framing become entangled with the sample itself. Although substantial progress has been made in recognizing richer affective signals (Cheng et al., 2024a), the same behavior cannot easily be re-observed under systematically changed conditions. It is therefore difficult to distinguish failures of affective understanding from failures of observation.

Interactive 3D simulation makes this missing dimension experimentally controllable (Savva et al., 2019). By instantiating affective behaviors as replayable events in executable worlds, the same behavior can be re-observed under different conditions, while an embodied agent can actively change the evidence available to it through action (Bajcsy, 1988; Yang et al., 2019). Simulation here is not merely a means of scaling data, but an experimental instrument for controlled and repeatable re-observation that is difficult to realize in the physical world.

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.

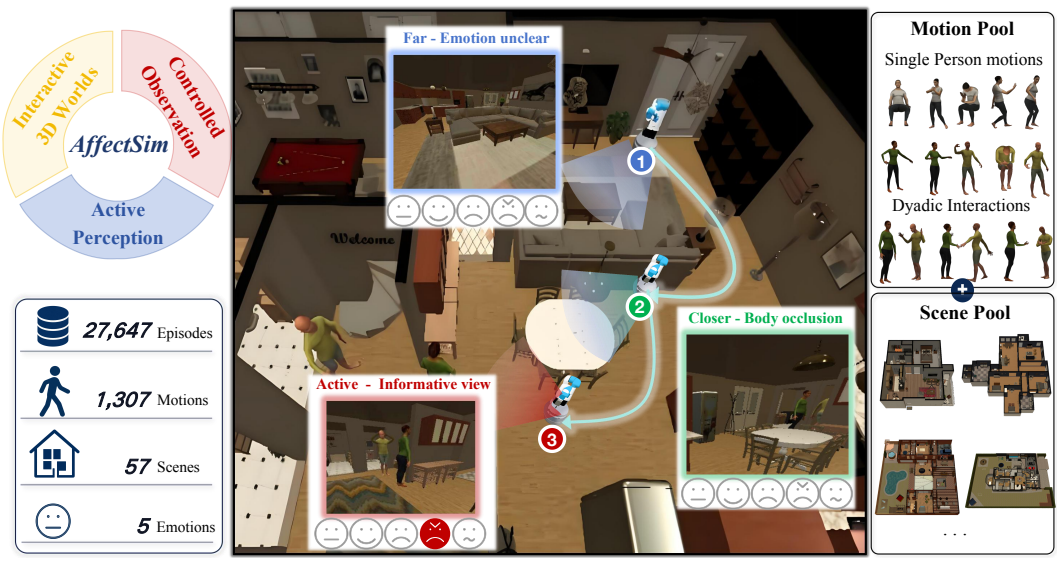


Figure 1: Overview of AffectSim. Reusable affective motions are instantiated in interactive 3D scenes under controllable observation conditions. In the illustrated episode, a distant view provides limited evidence, approaching remains insufficient under occlusion, and active repositioning reveals a more informative view.

To this end, we introduce AffectSim, a controllable interactive 3D benchmark for embodied affective perception. Its central design principle is to separate *what affective behavior is instantiated* from *how it is observed*. AffectSim contains 27,647 replayable episodes across 57 scenes, covering single-person and dyadic behaviors while independently varying viewpoint, distance, orientation, occlusion, and scene configuration. Matched passive and active observation protocols quantify the resulting *observation-quality gap* and how much of it can be recovered through embodied sensing.

Our contributions are threefold:

(1) A controllable benchmark for embodied affective perception. AffectSim enables matched re-observation of the same affective behavior under systematically varied observation conditions and agent actions.

(2) A systematic study of observation-driven affective perception. Experiments across 24 frozen recognizer configurations reveal a substantial observation-quality gap and show that active sensing can partially recover it.

(3) Reference solutions for active affective observation. We propose a training-free heuristic baseline and a learned closed-loop system with Evidence-Aware Observation Gate (EAOG), providing initial reference results for active evidence acquisition and its quality–cost trade-off.

2 Related Work

Affective perception from fixed observations. Affective computing has progressively expanded from facial and conversational cues to contextual, bodily, multimodal, and motion-level evidence (Poria et al., 2019; Kosti et al., 2017; Liu et al., 2021; Fourati & Pelachaud, 2014; Zhang et al., 2020). Recent foundation-model benchmarks further emphasize affective reasoning and explanation (Cheng et al., 2024a; Lian et al., 2025; Zhang et al., 2026). Despite increasingly rich evidence and stronger models, these settings share a common assumption: observations are collected and fixed before inference, leaving how affective evidence is acquired largely outside the evaluation loop.

Embodied affective perception and active sensing. Robot-centric datasets, E^3, OSMO, RobotEQ, and EmpathyAgent move affective understanding toward egocentric or embodied settings (Quiroz et al., 2022; Lin et al., 2024; Abdelfattah et al., 2026; Fang et al., 2026; Chen et al., 2025a). However, an egocentric viewpoint does not by itself make observation controllable: the physical trajectory that determines the available evidence is still largely predefined. Active perception in-

Table 1: Comparison of representative affective and embodied benchmarks. ✓ and ✗ indicate whether each capability is supported in an official evaluation setting.

Benchmark	Egocentric view	3D humans	Interactive environment	Controlled replay	Agent control
Passive affective benchmarks					
Emilya (Fourati & Pelachaud, 2014)	✗	✗	✗	✗	✗
KDAE (Zhang et al., 2020)	✗	✗	✗	✗	✗
DFEW (Jiang et al., 2020)	✗	✗	✗	✗	✗
iMiGUE (Liu et al., 2021)	✗	✗	✗	✗	✗
MER2024 (Lian et al., 2024)	✗	✗	✗	✗	✗
MERBench (Lian et al., 2026)	✗	✗	✗	✗	✗
Egocentric and social-affective benchmarks					
E^3 (Lin et al., 2024)	✓	✗	✗	✗	✗
EmpathyAgent (Chen et al., 2025a)	✓	✓	✗	✗	✗
OSMO (Abdelfattah et al., 2026)	✓	✗	✗	✗	✗
RobotEQ (Fang et al., 2026)	✓	✗	✗	✗	✗
AffectSim (ours)	✓	✓	✓	✓	✓

stead couples sensing with action so that an agent can change subsequent observations (Bajcsy, 1988; Yang et al., 2019; Fan et al., 2024). AffectSim brings this capability to affective perception by enabling the same affective behavior to be replayed under controlled observation conditions while allowing an embodied agent to actively acquire new evidence. We summarize the key differences among representative benchmarks in Table 1. A detailed comparison with existing benchmarks is provided in Section A.

3 AFFECTSIM BENCHMARK

AffectSim is built around a simple principle: *affective behavior and its observation conditions should be independently controllable*. As illustrated in Figure 2, emotion-labeled performances are converted into reusable motion assets and composed with 3D scenes, human placements, and observation challenges to form executable episodes.

The benchmark is implemented in Habitat (Savva et al., 2019) with HSSD indoor scenes (Khanna et al., 2024). We represent an episode as

$$e = \mathcal{C}(m, S, p_h, x_0), \tag{1}$$

where m is an affective motion, S a scene, p_h the human placement and target identity, and x_0 the agent initialization. The emotion label is inherited from m, while distance, orientation, occlusion, scene geometry, and viewpoint are determined by the embodied configuration. Thus, the same underlying behavior can be re-observed under different conditions rather than being tied to a pre-rendered video.

3.1 AFFECTIVE MOTION ASSETS

The motion library combines KDAE (Zhang et al., 2020), Emilya (Fourati & Pelachaud, 2014), and a newly captured dyadic pilot dataset, covering scenario-based, everyday-action, and interpersonal affective behaviors. We harmonize their annotations into five shared classes: *angry*, *fearful*, *happy*, *neutral*, and *sad*, excluding categories without clear cross-dataset correspondence. Motions that can be faithfully instantiated in simulation are retained and retargeted to a common avatar while preserving their source labels and provenance. A blind human validation on 500 randomly sampled retargeted motions achieves 77.81% Macro-F1 against the source emotion labels, supporting that affective semantics are largely preserved after retargeting. Detailed motion selection, retargeting, and validation procedures are provided in Appendix B.

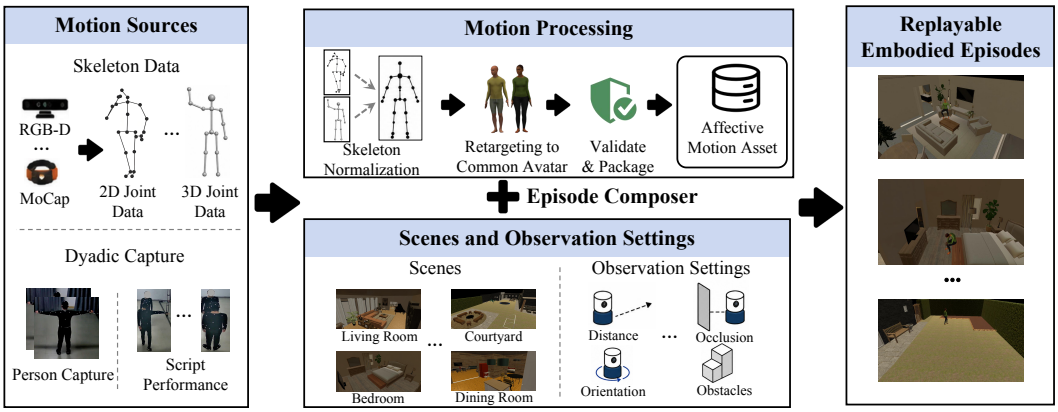


Figure 2: Construction pipeline of AffectSim. Emotion-labeled performances are converted into reusable motion assets and composed with scenes and controlled observation conditions to form replayable embodied episodes.

3.2 Embodied Episode Generation

The episode composer instantiates eq. (1) using two complementary construction modes according to whether the motion semantics depend on the surrounding environment.

Action-grounded composition. Motions such as sitting, door knocking, and walking are aligned with compatible scene geometry and navigable trajectories to ensure physically and semantically valid execution.

Observation-controlled composition. For motions without object-specific dependencies, the affective behavior is held fixed while the initial observation is varied. We construct four challenge families—*far-visible*, *rear-view*, *partial-occlusion*, and *obstacle-detour*—to stress target scale, viewing direction, body visibility, and navigational accessibility. The same abstraction also supports dyadic behavior.

Episode execution. The robot is initialized on navigable floor space and the affective motion is replayed cyclically during the episode. This factors out event-timing uncertainty and isolates the spatial active-perception problem: whether an agent can acquire a viewpoint that exposes discriminative bodily cues.

3.3 Validation, Scale, and Splits

Generated episodes undergo automated checks for configuration integrity, physical validity, reproducibility, and realization of the intended observation challenge. A pool of 15 reviewers further screens candidate episodes for physical plausibility, motion–scene compatibility, and rendering quality; only episodes judged reasonable are retained.

The final benchmark contains 27,647 episodes across 57 scenes, with 19,662/4,044/3,941 episodes in the train/validation/test splits. All variants derived from the same underlying performance are assigned to the same split to prevent motion leakage. Scenes are shared across splits, so the benchmark evaluates generalization to unseen performers and source motions within a common scene inventory rather than to unseen environments. Detailed split statistics, validation criteria, and initialization procedures are provided in Section B.

4 Evaluation Protocol

4.1 Matched Observation Settings

All paired settings preserve the scene, target motion, motion phase, emotion label, rendering parameters, and robot initialization. Within each experiment, they also share the same temporal budget and recognition interface; only the observation trajectory changes.

P-Init (Initial View). The camera remains fixed at the agent's initial pose x_0, measuring the evidence available before any embodied action. **P-Ref (Reference View).** A privileged reference observer maintains a reproducible informative view according to body visibility, alignment, distance, framing, and scene clearance. P-Ref is a diagnostic reference rather than an oracle or upper bound. **P-Mon (Monitor View).** A fixed room-level camera provides an external monitoring view and serves as an additional passive diagnostic. **A-Obs (Active View).** A-Obs starts from the same robot and human state as P-Init, but allows the observation policy to change the robot viewpoint.

4.2 Evaluation Metrics

Recognition quality. We use Macro-F1 as the primary five-class recognition metric and additionally report accuracy and probability-based recognition metrics where appropriate.

Observation benefit. For a higher-is-better metric M and active policy π, the paired gain is $\Delta_{\text{gain}}(M;\pi) = M_\pi - M_{\text{P-Init}}$. To quantify how much of the observation-quality gap is recovered, we define

$$\text{OGR}(M;\pi) = \frac{M_\pi - M_{\text{P-Init}}}{M_{\text{P-Ref}} - M_{\text{P-Init}}}, \tag{2}$$

when $M_{\text{P-Ref}} > M_{\text{P-Init}}$. Because P-Ref is a reproducible reference rather than an oracle, OGR is not clipped to $[0, 1]$.

Recovery and acquisition cost. Recovery Rate measures how often active observation corrects an initially incorrect prediction. We additionally report path length and Emotion-SPL (E-SPL) as navigation-aware acquisition diagnostics following Anderson et al. (2018). Formal definitions and supplementary execution diagnostics are provided in Section C.

5 Active Affective Perception

We formulate embodied affective perception as a closed-loop active sensing problem, as illustrated in Figure 3. At each step, the agent observes the target, adjusts its viewpoint, and updates its accumulated affective evidence. The observation policy determines *where to observe*, while the Evidence-Aware Observation Gate (EAOG) evaluates the accumulated affective evidence and determines whether continued observation remains worthwhile. Thus, viewpoint control and affective evidence assessment proceed concurrently throughout interaction.

5.1 Closed-Loop Active Observation

Embodied observation policy. We adapt Uni-NaVid (Zhang et al., 2025) with a movement head over five actions: Forward, Backward, Left, Right, and Wait. At each step, the policy receives recent egocentric RGB observations, an instruction to maintain an informative view of the target person, and its executed-action history, and predicts the next movement action.

Evidence-aware observation termination. In parallel with viewpoint control, EAOG continuously evaluates the affective evidence acquired so far. At each decision time $t \in \mathcal{T}$, it extracts spatiotemporal features from the current RGB window $\mathcal{W}_t$ and updates emotion-conditioned evidence memories:

$$\mathbf{z}_{t,c} = \text{Attend}(\mathbf{q}_c, \text{Enc}(\mathcal{W}_t)), \qquad \mathbf{e}_{t,c} = \text{GRU}(\mathbf{z}_{t,c}, \mathbf{e}_{t-1,c}). \tag{3}$$

The accumulated evidence states produce a termination probability:

$$h_t = \sigma(g_{\text{term}}([\mathbf{e}_{t,1}; \ldots; \mathbf{e}_{t,C}])) \,. \tag{4}$$

Here, h_t measures whether further sensing is still worthwhile rather than emotion confidence itself. If the accumulated evidence remains insufficient, the agent continues moving and observing; otherwise, active observation terminates. The resulting system therefore follows a continuous *move–observe–update–assess* loop.

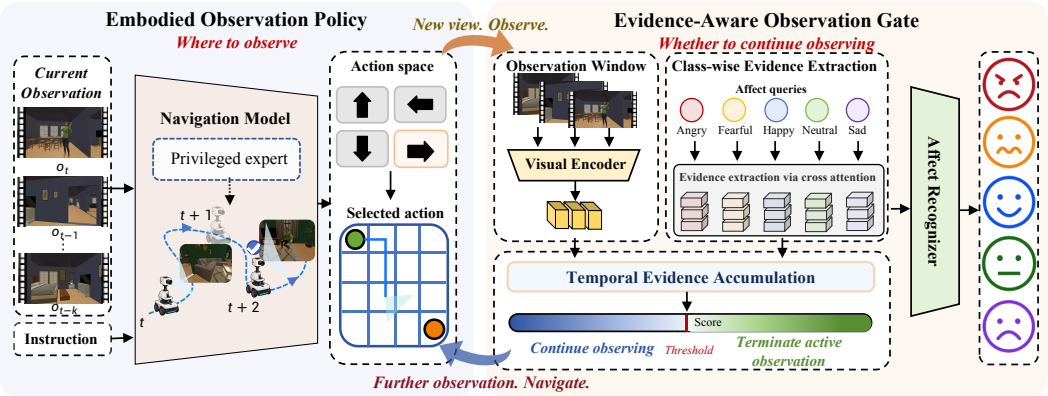


Figure 3: Overview of our closed-loop active affective perception framework. As the agent moves and observes, the observation policy controls viewpoint acquisition while EAOG continuously accumulates affective evidence and assesses whether further observation is beneficial.

5.2 Learning the Active Observation System

Learning where to observe. We train the observation policy through behavior cloning from a privileged teacher. The teacher uses simulator-side information, including scene geometry, navigation paths, target motion, and body visibility, to generate actions that improve affective evidence acquisition. The student learns these actions using only observations available at inference.

Learning when to terminate observation. EAOG is trained from observation trajectories while the learned observation policy and affect recognizer remain frozen. A frozen affect recognizer evaluates each candidate window and provides a recognition risk R_t based on the Half-Brier score. We combine recognition risk with observation delay:

$$C_t = R_t + \lambda(t - t_{\min}), \qquad d_t = \frac{\exp(-C_t/\tau)}{\sum_j \exp(-C_j/\tau)}. \tag{5}$$

The soft target d_t supervises when additional observation is no longer worthwhile. Emotion classification and quality prediction provide auxiliary supervision:

$$\mathcal{L}_{\text{EAOG}} = \mathcal{L}_{\text{term}} + \mathcal{L}_{\text{quality}} + \mathcal{L}_{\text{emotion}}. \tag{6}$$

Detailed scoring, supervision, and optimization are provided in Section E.3.

6 Experiments

6.1 Experimental Setup

We evaluate all 3,941 test episodes under two complementary settings. In the *zero-shot setting*, 24 frozen affect recognizers (19 open-source and 5 closed-source) compare P-Init, P-Mon, P-Ref, and a training-free heuristic active baseline. The baseline uses a frozen planner to locate the target, followed by a depth- and location-based controller to maintain an informative view, without emotion supervision or recognizer feedback. If no valid new observation is acquired, it falls back to the paired P-Init view.

In the *learned setting*, we evaluate EAOG with a behavior-cloned observation policy and an AffectSim-adapted frozen VideoLLaMA2.1-7B-AV recognizer. Controlled comparisons fix the navigator, recognizer, and candidate observations, varying only the observation-gating strategy. Implementation and inference details are provided in Sections D and E.

6.2 Q1: Does Observation Quality Matter?

We isolate the effect of observation quality by holding the affective motion, scene, motion phase, temporal budget, and recognizer interface fixed for each matched episode–recognizer pair, while

Table 2: Mean test Macro-F1 (%) across 24 frozen affect recognizers under matched observation conditions. P-Init, P-Mon, and P-Ref denote passive observations; Heuristic denotes the training-free active-observation baseline, with P-Init fallback when no valid new observation is acquired. Full per-recognizer results are reported in Appendix Table D.

Recognizer group	# Models	Examples	P-Mon	P-Ref	P-Init	Heur.	Heur.−P-Init
Open-source	19	Qwen, AffectGPT	10.31	15.56	9.89	11.70	+1.81
Closed-source	5	GPT, Gemini, Claude	24.44	30.78	22.61	24.26	+1.65
Overall	24	–	13.26	18.73	12.54	14.32	+1.78

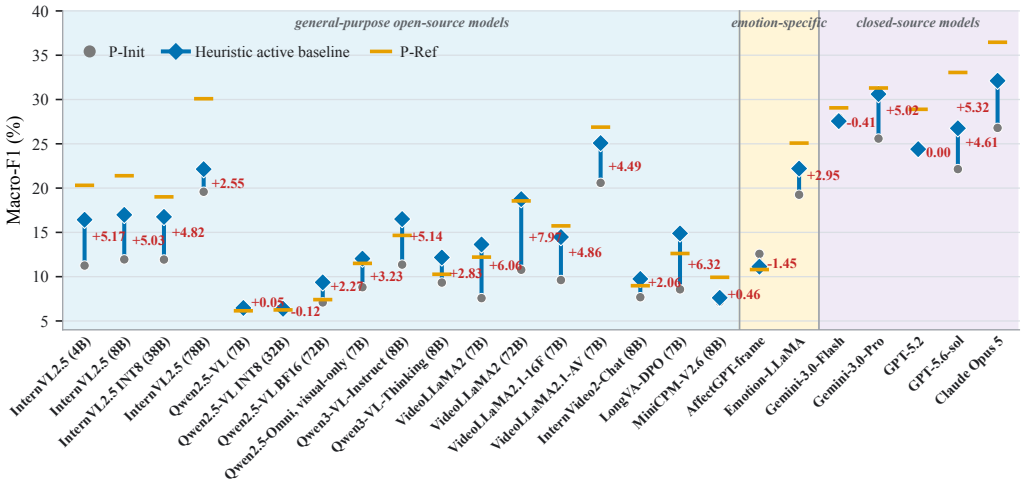


Figure 4: Paired Macro-F1 (%) of 24 frozen affect recognizers on the 1,881 test episodes where the heuristic baseline successfully acquires a valid new observation. Red labels show the paired gain over P-Init. This subset isolates the quality of newly acquired observations from acquisition failures.

varying only the observation condition. We compare the initial egocentric view (P-Init), fixed room-monitor view (P-Mon), and privileged reference view (P-Ref).

As shown in Table 2, P-Ref substantially improves recognition over P-Init across both open- and closed-source recognizers. Mean Macro-F1 increases from 9.89% to 15.56% for the 19 open-source configurations and from 22.61% to 30.78% for the five closed-source configurations; P-Ref achieves the best passive-view performance for 23 of 24 configurations. This establishes a substantial *observation-quality gap*: recognition performance depends not only on the recognizer, but also on the evidence made available to it.

The gain is not explained simply by replacing the egocentric view with an external camera. P-Mon reaches only 10.31% and 24.44% Macro-F1 for the open- and closed-source groups, respectively, remaining much closer to P-Init than to P-Ref. Thus, informative observation depends on factors such as visibility, distance, orientation, and framing rather than camera placement alone. P-Ref is a reproducible diagnostic reference rather than an oracle or upper bound. Complete per-recognizer results for all passive observation conditions are reported in Table D.

6.3 Q2: Does Active Observation Improve Recognition?

We next ask whether actively changing the viewpoint can improve affect recognition. We evaluate the training-free heuristic baseline, which uses neither emotion supervision nor recognizer feedback.

Broad zero-shot gains. On the complete test set, the heuristic baseline improves Macro-F1 for 21 of 24 frozen recognizer configurations (Table 2). Mean Macro-F1 increases from 9.89% to 11.70% for the 19 open-source recognizers and from 22.61% to 24.26% for the five closed-source recognizers.

Table 3: Controlled comparison of observation strategies on test episodes using the same AffectSim-adapted and frozen VideoLLaMA2.1-7B-AV recognizer. P-Init is the reference for OGR.

Method	Observation strategy	F1↑	OGR↑	E-SPL@3↑
P-Init	Fixed initial view	32.74	–	33.80
Heuristic	Hand-designed active observation	34.52	8.66	15.89
EAOG	Learned active observation	**45.47**	**61.92**	**34.57**

Table 4: Recognition and acquisition efficiency under different observation-termination strategies. The navigator and recognizer are fixed; only the termination strategy differs.

Method	F1↑	OGR↑	E-SPL@3↑
Horizon mid. (24 s)	43.02	50.00	32.04
Candidate mid. (30 s)	43.94	54.47	31.89
Final (48 s)	43.93	54.43	29.79
Uniform random	44.27	56.08	32.18
EAOG	**45.47**	**61.92**	**34.57**

Table 5: Ablation of EAOG training objectives. Each variant removes one training signal while keeping the remaining training and evaluation settings unchanged.

Method	F1↑	OGR↑	E-SPL@3↑
w/o emotion loss	42.65	48.20	31.83
w/o quality loss	44.65	57.93	32.52
w/o waiting cost	42.10	45.53	31.64
EAOG	**45.47**	**61.92**	**34.57**

These gains recover 32.0% and 20.1% of the corresponding P-Ref–P-Init observation-quality gaps, respectively. Thus, active viewpoint acquisition improves affect recognition even without emotion-specific supervision. Complete per-recognizer results, Recovery Rate, and path-aware E-SPL results are reported in Tables D and G.

Successful acquisition. The complete-set result also reflects acquisition failures. The heuristic baseline successfully acquires a valid new observation in 1,881 of the 3,941 test episodes. On this successful-acquisition subset (Figure 4), the gains over paired P-Init increase to 3.40 and 2.91 percentage points for the open- and closed-source groups, respectively, with improvements for 20 of the 24 recognizer configurations. This shows that successfully acquired viewpoints generally provide more useful affective evidence, while reliable viewpoint acquisition remains an important bottleneck. Detailed subset results are reported in Table F.

Learned vs. heuristic observation. We further compare heuristic and learned active observation using the same AffectSim-adapted and frozen VideoLLaMA2.1-7B-AV recognizer. As shown in Table 3, P-Init achieves 32.74% Macro-F1, P-Ref achieves 53.3%, the heuristic baseline reaches 34.52%, and EAOG reaches 45.47%. This controlled comparison shows that learned active observation provides a substantial improvement beyond the hand-designed heuristic baseline. Q3 further isolates the contribution of evidence-aware termination by fixing the learned navigator and recognizer and varying only the termination strategy.

6.4 Q3: Can Evidence-Aware Observation Gate Improve Recognition within a Limited Observation Budget?

Adaptive vs. non-adaptive termination. We compare EAOG with four non-adaptive strategies: terminating at the midpoint of the full observation horizon (Horizon midpoint, 24 s), at the midpoint of the valid candidate interval (Candidate midpoint, 30 s), at the final observation (48 s), or at a uniformly sampled candidate time. These controls represent simple alternatives that use the same observations but do not adapt the termination time to the acquired affective evidence.

As shown in Table 4, EAOG achieves the best performance across all three metrics, with 45.47% Macro-F1, 61.92% OGR, and 34.57 E-SPL@3. It outperforms the best fixed-time strategy by 1.53 percentage points and uniform random termination by 1.20 points in Macro-F1. EAOG terminates at 33.72 s on average, 14.28 s earlier than the 48-s endpoint. These results show that the most useful affective evidence does not necessarily occur at the end of observation, and that adapting the termination time to accumulated evidence improves both recognition quality and acquisition efficiency. Detailed stopping-time and recognition statistics are reported in Section F.4.

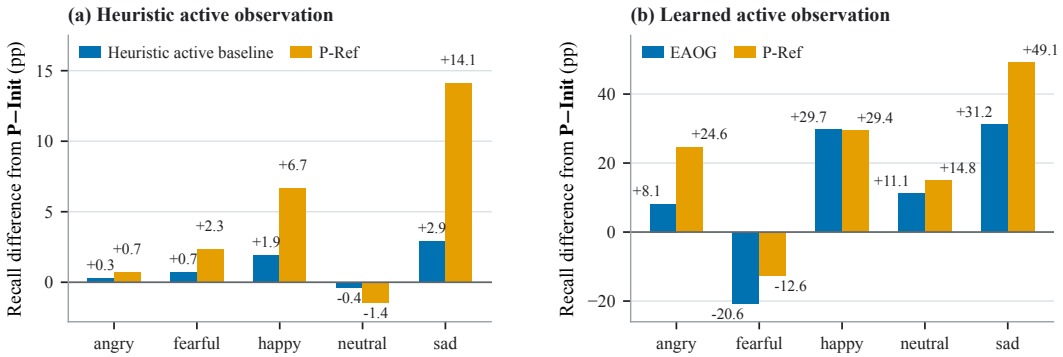


Figure 5: Class-wise recall over the paired P-Init observation. (a) Heuristic-baseline averaged over all 24 frozen recognizer configurations. (b) EAOG under the controlled recognizer setting. P-Ref is shown as a diagnostic reference. All values are percentage-point (pp) gains.

EAOG objective ablation. We further ablate the emotion loss, quality loss, and waiting cost while keeping the remaining training and evaluation settings unchanged (Table 5). Removing these components reduces Macro-F1 from 45.47% to 42.65%, 44.65%, and 42.10%, respectively. In particular, removing the waiting cost causes the largest degradation, indicating that explicitly balancing recognition quality against delayed observation is important for effective termination. The full EAOG also achieves the highest OGR and E-SPL@3. Additional accuracy, Half-Brier, stopping-time, and threshold-selection results are provided in Table E and Section E.5.

6.5 Q4: Where Does Active Observation Help, and What Does It Cost?

Class-wise variation. As shown in Figure 5, active observation does not benefit all emotions equally. For the heuristic baseline, the largest average recall gains occur for sad (+2.9 pp) and happy (+1.9 pp), while neutral slightly decreases (-0.4 pp). The P-Ref results further show that the observation-quality gap itself varies substantially across emotion categories. Under the controlled EAOG setting, the class-wise changes are larger: recall improves for happy (+29.7 pp), sad (+31.2 pp), neutral (+11.1 pp), and angry (+8.1 pp), but decreases for fearful (-20.6 pp). Thus, learned active observation yields substantial overall gains but also introduces a pronounced class-dependent trade-off. More results are reported in the appendix.

Acquisition cost. Improved observations are not free to acquire. For the heuristic baseline, 47.73% of test episodes reach stable handoff, while 52.27% exhaust the action budget beforehand. The robot travels 23.23 m per episode on average, and 97.4% of locomotion occurs before handoff. This indicates that most of the acquisition cost lies in locating and approaching the target rather than maintaining the final observation. Together with the efficiency results in Q3, these findings identify viewpoint acquisition itself as a major bottleneck for active affective perception. More execution statistics, per-model Recovery Rate and path-aware E-SPL are reported in Tables G and H.

7 Conclusion

We introduce AffectSim, a controllable interactive 3D benchmark for embodied affective perception, together with a training-free heuristic baseline and the learned EAOG method. Our results establish observation quality as a key factor in affect recognition and demonstrate the value of active observation. EAOG improves recognition while reducing observation time relative to the full-horizon baseline. More importantly, AffectSim uses simulation as an experimental instrument rather than merely a means of scaling data, enabling controlled and repeatable re-observation that is difficult to realize in the physical world. It provides a foundation for studying affective perception as an embodied sensing problem. **Limitations and Future Directions:** AffectSim currently focuses on body-motion cues and predefined human behaviors. Future work can incorporate facial, speech, and other multimodal cues, more realistic avatars and environments, reactive virtual humans, and real-robot evaluation toward richer interactive affective intelligence.

ETHICS STATEMENT

Emotion-perception technologies raise concerns about surveillance, manipulation, psychological profiling, privacy, and culturally dependent labels. The motion-capture protocol follows procedures for informed consent, compensation, data governance, withdrawal, and data use. Released avatars do not reproduce participants' real facial appearance, reducing direct identity exposure relative to in-the-wild video. AffectSim is intended for research on embodied perception, assistive robotics, and human-centered AI, not high-stakes decision making or surveillance. Release documentation will describe label subjectivity and cultural limitations, intended uses, and procedures for reporting potential misuse.

AI USE STATEMENT

Generative AI tools were used to assist with language compression and LaTeX formatting. The authors remain responsible for verifying all technical content, citations, claims, and reported results.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide our code in an anonymous repository at `https://anonymous.4open.science/r/AffectSim-8B44/README.md`. Section 3 describes the episode construction procedure. Section 4 specifies the paired observation protocol. The appendix provides complete metric definitions, data split specifications, execution details, and results for each configuration.

REFERENCES

Mohamed Abdelfattah, Bugra Tekin, Fadime Sener, Necati Cihan Camgoz, Eric Sauser, Shugao Ma, Alexandre Alahi, and Edoardo Remelli. OSMO: Open-vocabulary Self-eMOtion tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1737–1748, 2026.

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):996–1005, 1988. doi: 10.1109/5.5968.

Uttaran Bhattacharya, Trisha Mittal, Rohan Chandra, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha. STEP: Spatial temporal graph convolutional networks for emotion perception from gaits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(2):1342–1350, 2020. doi: 10.1609/aaai.v34i02.5490.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008. doi: 10.1007/s10579-008-9076-6.

Xinyan Chen, Jiaxin Ge, Hongming Dai, Qiang Zhou, Qiuxuan Feng, Jingtong Hu, Yizhou Wang, Jiaming Liu, and Shanghang Zhang. EmpathyAgent: Can embodied agents conduct empathetic actions? *arXiv preprint arXiv:2503.16545*, 2025a.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2025b.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G Hauptmann. Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 110805–110853, 2024a.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024b.

Lei Fan, Jianxiong Zhou, Xiaoying Xing, and Ying Wu. Active open-vocabulary recognition: Let intelligent moving mitigate CLIP limitations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16394–16403, 2024. doi: 10.1109/CVPR52733.2024.01551.

Kuofei Fang, Zheng Lian, Xinyi Che, Haomin Ouyang, Shufan Zhang, Xuehao Wang, Qi Liu, Liyi Liu, Chenqi Zhang, Wenxi Cai, et al. RobotEQ: Towards social proactive intelligence in embodied agents. *arXiv preprint arXiv:2605.06234*, 2026.

Nesrine Fourati and Catherine Pelachaud. Emilya: Emotional body expression in daily actions database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3486–3493, 2014.

He Hu, Lianzhong You, Hongbo Xu, Qianning Wang, Fei Richard Yu, Fei Ma, Zebang Cheng, Zheng Lian, Yucheng Zhou, and Laizhong Cui. EmoBench-M: Benchmarking emotional intelligence for multimodal large language models. *arXiv preprint arXiv:2502.04424*, 2025.

Yibo Huang, Hongqian Wen, Linbo Qing, Rulong Jin, and Leiming Xiao. Emotion recognition based on body and context fusion in the wild. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3602–3610, 2021. doi: 10.1109/ICCVW54120.2021.00403.

Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. DFEW: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2881–2889, 2020. doi: 10.1145/3394171.3413620.

Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat Synthetic Scenes Dataset (HSSD-200): An analysis of 3D scene scale and realism tradeoffs for ObjectGoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16384–16393, 2024. doi: 10.1109/CVPR52733.2024.01550.

Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. EMOTIC: Emotions in context dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2309–2317. IEEE, 2017. doi: 10.1109/CVPRW.2017.285.

Deng Li, Bohao Xing, Xin Liu, Baiqiang Xia, Bihan Wen, and Heikki Kälviäinen. DEEMO: Deidentity multimodal emotion recognition and reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5707–5716. ACM, 2025. doi: 10.1145/3746027.3755411.

Zheng Lian, Haiyang Sun, Licai Sun, Hao Gu, Zhuofan Wen, Siyuan Zhang, Shun Chen, Mingyu Xu, Ke Xu, Kang Chen, et al. Explainable multimodal emotion recognition. *arXiv preprint arXiv:2306.15401*, 2023.

Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, et al. MER 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pp. 41–48. ACM, 2024. doi: 10.1145/3689092.3689959.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566*, 2025.

Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. MERBench: A unified evaluation benchmark for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(5):5793–5810, 2026. doi: 10.1109/TPAMI.2026.3653457.

Wang Lin, Yueying Feng, Wenkang Han, Tao Jin, Zhou Zhao, Fei Wu, Chang Yao, and Jingyuan Chen. E^3: Exploring embodied emotion through a large-scale egocentric video dataset. In *Advances in Neural Information Processing Systems*, volume 37, pp. 118182–118197, 2024. doi: 10.52202/079017-3753.

Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pp. 612–630. Springer, 2022a. doi: 10.1007/978-3-031-20071-7_36.

Shifeng Liu, Zhengye Zhang, Sirui Zhao, Xinglong Mao, Zhehan Kan, Zhixiang Wei, Shiwei Wu, Chaoyou Fu, Tong Xu, and Enhong Chen. ActFER: Agentic facial expression recognition via active tool-augmented visual reasoning. *arXiv preprint arXiv:2604.08990*, 2026.

Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. iMiGUE: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10631–10642, June 2021.

Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. MAFW: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 24–32, 2022b. doi: 10.1145/3503161.3548190.

Haifeng Lu, Xiping Hu, and Bin Hu. See your emotion from gait using unlabeled skeleton data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1826–1834, June 2023. doi: 10.1609/aaai.v37i2.25272.

Haifeng Lu, Jiuyi Chen, Feng Liang, Mingkui Tan, Runhao Zeng, and Xiping Hu. Understanding emotional body expressions via large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):1447–1455, April 2025a. doi: 10.1609/aaai.v39i2.32135.

Haifeng Lu, Jiuyi Chen, Zhen Zhang, Ruida Liu, Runhao Zeng, and Xiping Hu. Emotion recognition from skeleton data: A comprehensive survey. *arXiv preprint arXiv:2507.18026*, 2025b.

Yu Luo, Jianbo Ye, Reginald B. Adams, Jr., Jia Li, Michelle G. Newman, and James Z. Wang. ARBEE: Towards automated recognition of bodily expression of emotion in the wild. *International Journal of Computer Vision*, 128(1):1–25, 2020. doi: 10.1007/s11263-019-01215-y.

Venkatraman Narayanan, Bala Murali Manoghar, Vishnu Sashank Dorbala, Dinesh Manocha, and Aniket Bera. ProxEmo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 8200–8207, 2020.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1050.

Marco Quiroz, Raquel Patiño, Jose Díaz-Amado, and Yudith Cardinale. Dataset to detect emotions from a robot-centric perspective. In *Workshops at the 18th International Conference on Intelligent Environments (IE2022)*, Ambient Intelligence and Smart Environments. IOS Press, 2022. doi: 10.3233/AISE220019.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9339–9347, 2019.

Zixuan Shangguan, Yanjie Dong, Song Guo, Victor C. M. Leung, M. Jamal Deen, and Xiping Hu. Facial expression analysis and its potentials in IoT systems: A contemporary survey. *ACM Computing Surveys*, 58(2):43, 2026. doi: 10.1145/3737456.

Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. FERV39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20922–20931, 2022. doi: 10.1109/CVPR52688.2022.02025.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. InternVideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision – ECCV 2024*, pp. 396–416. Springer, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Jianwei Yang, Zhile Ren, Mingze Xu, Xinlei Chen, David J. Crandall, Devi Parikh, and Dhruv Batra. Embodied amodal recognition: Learning to move to perceive objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2040–2050, 2019.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Yunfei Yin, Li Jing, Faliang Huang, Guangchao Yang, and Zhuowei Wang. MSA-GCN: Multi-scale adaptive graph convolution network for gait emotion recognition. *Pattern Recognition*, 147: 110117, 2024. doi: 10.1016/j.patcog.2023.110117.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, et al. MME-Emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. In *International Conference on Learning Representations*, 2026.

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-NaVid: A video-based vision-language-action model for unifying embodied navigation tasks. In *Robotics: Science and Systems XXI*, 2025. doi: 10.15607/RSS.2025.XXI.013.

Mingming Zhang, Lu Yu, Keye Zhang, Bixuan Du, Bin Zhan, Shaohua Chen, Xiuhao Jiang, Shuai Guo, Jiafeng Zhao, Yang Wang, et al. Kinematic dataset of actors expressing emotions. *Scientific Data*, 7(1):292, 2020.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

APPENDIX CONTENTS

Section A: Detailed Benchmark Comparison 15
Section B: Supplementary Benchmark Construction Details 16
Section C: Supplementary Metric Definitions 18
Section D: Training and Implementation Details 18
Section E: EAOG Architecture and Training Details 19
Section F: Additional Experimental Results 23
Section G: Limitations, Ethics, and Future Work 28

A Detailed Benchmark Comparison

A.1 Expanded Related-Work Discussion

Table A: Comparison of representative affective and embodied benchmarks. ✓ and ✗ indicate whether each capability is supported in an official evaluation setting. Active perception refers to acquiring new observations through agent actions.

Benchmark	Active perception	Egocentric view	3D humans	Interactive environment	Controlled replay	Agent control
Passive affective benchmarks						
DFEW (Jiang et al., 2020)	✗	✗	✗	✗	✗	✗
MAFW (Liu et al., 2022b)	✗	✗	✗	✗	✗	✗
MER2024 (Lian et al., 2024)	✗	✗	✗	✗	✗	✗
Emilya (Fourati & Pelachaud, 2014)	✗	✗	✗	✗	✗	✗
KDAE (Zhang et al., 2020)	✗	✗	✗	✗	✗	✗
BoLD (Luo et al., 2020)	✗	✗	✗	✗	✗	✗
iMiGUE (Liu et al., 2021)	✗	✗	✗	✗	✗	✗
IEMOCAP (Busso et al., 2008)	✗	✗	✗	✗	✗	✗
MELD (Poria et al., 2019)	✗	✗	✗	✗	✗	✗
EMOTIC (Kosti et al., 2017)	✗	✗	✗	✗	✗	✗
MERBench (Lian et al., 2026)	✗	✗	✗	✗	✗	✗
DEEMO (Li et al., 2025)	✗	✗	✗	✗	✗	✗
MME-Emotion (Zhang et al., 2026)	✗	✗	✗	✗	✗	✗
Egocentric and social-affective benchmarks						
Robot-centric Emotion Dataset (Quiroz et al., 2022)	✗	✓	✓	✗	✗	✗
E^3 (Lin et al., 2024)	✗	✓	✗	✗	✗	✗
OSMO (Abdelfattah et al., 2026)	✗	✓	✗	✗	✗	✗
RobotEQ (Fang et al., 2026)	✗	✓	✗	✗	✗	✗
EmpathyAgent (Chen et al., 2025a)	✗	✓	✓	✗	✗	✗
AffectSim (ours)	✓	✓	✓	✓	✓	✓

We retain the fuller discussion here because it explains more precisely which part of the affective-perception pipeline is placed inside the evaluation loop.

Emotion understanding from fixed observations. Most affective computing benchmarks evaluate models from observations fixed before inference. IEMOCAP and MELD study emotion recognition from prerecorded multimodal conversations (Busso et al., 2008; Poria et al., 2019); DFEW and MAFW focus on dynamic facial and multimodal emotion recognition in the wild (Jiang et al., 2020; Liu et al., 2022b); and MER2024 and MERBench provide broader multimodal evaluation protocols (Lian et al., 2024; 2026). Other benchmarks extend the supplied evidence beyond facial appearance. EMOTIC incorporates scene context, BoLD and iMiGUE emphasize bodily or non-facial behavior, and structured motion datasets such as Emilya, KDAE, STEP, and BEAT provide explicit motion-level representations (Kosti et al., 2017; Luo et al., 2020; Liu et al., 2021; Fourati & Pelachaud, 2014; Zhang et al., 2020; Bhattacharya et al., 2020; Liu et al., 2022a). These resources support facial, audiovisual, contextual, and body-based emotion recognition (Yin et al., 2024; Shangguan et al., 2026; Lu et al., 2023; 2025b;a; Huang et al., 2021).

Recent multimodal foundation models extend affective understanding from categorical recognition toward explanation and reasoning. Emotion-LLaMA and EMER model multimodal emotional evidence and its interpretation, while AffectGPT, EmoBench-M, and MME-Emotion broaden evalu-

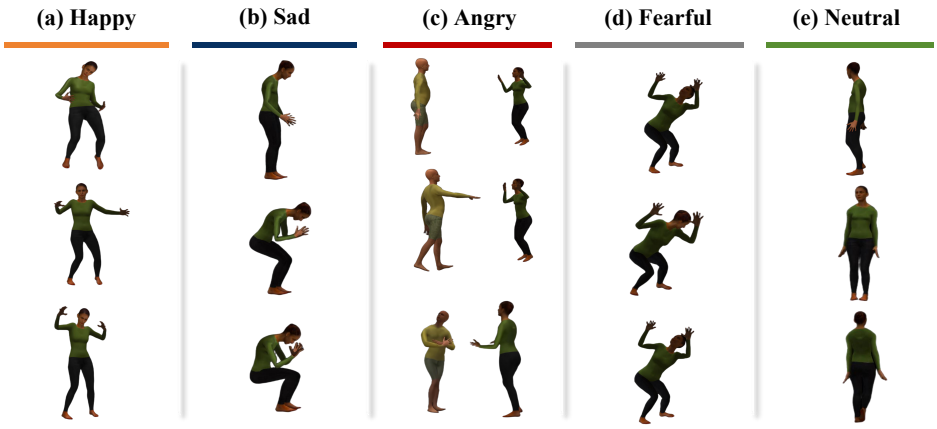


Figure A: Representative AffectSim motion assets across five emotion categories, covering both single-person and dyadic interactions.

ation toward free-form and reasoning-oriented affective understanding (Cheng et al., 2024a; Lian et al., 2023; 2025; Hu et al., 2025; Zhang et al., 2026). Despite richer evidence and stronger reasoning, these settings retain a common assumption: the observation itself is predetermined. The evaluated model can recognize, fuse, or reason over supplied evidence, but cannot physically act to improve it or re-observe the same affective behavior under controlled conditions. AffectSim studies this upstream observation process.

Embodied affective perception and social intelligence. Recent work has moved affective computing toward robot-centric and egocentric settings. The Robot-centric Emotion Dataset captures affective behavior during human–robot interaction, while E^3 and OSMO provide first-person audio-visual observations in everyday or social scenarios (Quiroz et al., 2022; Lin et al., 2024; Abdelfattah et al., 2026). RobotEQ studies emotional intelligence in embodied scenarios using egocentric observations and spatial grounding (Fang et al., 2026), while ActFER adaptively selects facial regions for recognition (Liu et al., 2026). These works bring affective perception closer to the observer, but the physical observation trajectory remains predetermined: the evaluated system cannot reposition itself to alter subsequent visual evidence or systematically re-observe the same affective event.

Beyond perception, systems such as EmpathyAgent integrate affective perception, memory, reasoning, and action generation for empathetic interaction (Chen et al., 2025a). Their focus is primarily on how an embodied agent interprets supplied affective evidence and acts upon it, rather than on how that evidence should be physically acquired. In contrast, AffectSim places the observation process itself inside the evaluation loop. Its executable 3D episodes support controlled re-observation of the same affective behavior, and an agent's actions directly change the visual evidence subsequently available for emotion recognition.

B Supplementary Benchmark Construction Details

The main construction pipeline, quality-control criteria, and split statistics appear in section 3. This appendix provides motion examples and additional implementation checks.

B.1 Motion Retargeting Checks

The resulting motions are screened for skeletal distortions, implausible joint configurations, root-motion discontinuities, foot sliding, and visible body penetration.

Table B: Group-disjoint benchmark splits of AffectSim. All episodes generated from the same source performance remain in the same split.

Split	Action-grounded		Observation-controlled		Dyadic pilot		Total
	Actors	Episodes	Motions	Episodes	Motions	Episodes	Episodes
Train	8	11,999	669	6,643	34	1,020	19,662
Validation	2	3,003	84	831	7	210	4,044
Test	2	2,899	84	832	7	210	3,941
Total	12	17,901	837	8,306	48	1,440	27,647

Table C: Class-wise results of manual affective-semantic validation on the sampled retargeted motions.

Emotion	Precision (%)	Recall (%)	F1 (%)
Angry	73.68	78.65	76.09
Fearful	87.38	79.65	83.33
Happy	78.43	68.97	73.39
Sad	76.25	74.39	75.31
Neutral	74.17	89.00	80.91
Macro avg.	77.98	78.13	77.81

B.2 Episode Initialization and Integrity Checks

For action-grounded episodes, the robot and target occupy the same accessible local environment with the target initially visible; for observation-controlled episodes, x_0 is selected according to the assigned challenge family.

Scene and motion references must be resolvable, episode identifiers unique, metadata consistent with the release manifest, and robot initial states navigable. Motion-specific checks verify support alignment, hand–door geometry, standing-to-seated transitions, and collision-free walking trajectories.

B.3 Affective-Semantic Validation after Retargeting

To assess whether motion retargeting preserves the affective semantics of the source performances, we randomly sampled 500 retargeted motion assets for manual evaluation. Three reviewers independently inspected the rendered motions and assigned one of the five AffectSim emotion labels: angry, fearful, happy, sad, or neutral. The final label for each motion was determined by majority vote.

Compared with the source emotion labels, the retargeted motions achieve an overall accuracy of 78% and a Macro-F1 of 77.81%. Table C reports the class-wise precision, recall, and F1 scores. These results indicate that the affective semantics are largely preserved after retargeting.

B.4 Simulator Interface and Extensibility

The benchmark can expose egocentric RGB-D observations and proprioceptive state. Because actions alter future observations, one episode can support matched initial, reference, monitor, and actively acquired views. The asset–episode separation also makes the benchmark extensible: new affective motions, performers, dyadic behaviors, generated motions, avatars, scenes, and observation challenges can be added without redefining the episode abstraction or evaluation protocol.

C SUPPLEMENTARY METRIC DEFINITIONS

The matched observation settings, study-specific budget, and primary metric definitions appear in section 4. The following formulas specify the supplementary recovery and efficiency measures.

Episode-level recovery For paired P-Init and active predictions, Recovery Rate is

$$\text{Recovery Rate}(\pi) = \frac{\sum_i \mathbf{1}[\hat{y}_i^{\text{P-Init}} \neq y_i \wedge \hat{y}_i^{\pi} = y_i]}{\sum_i \mathbf{1}[\hat{y}_i^{\text{P-Init}} \neq y_i]}. \tag{7}$$

Acquisition efficiency At target-distance reference r,

$$\text{E-SPL@}r = \frac{1}{N_r} \sum_{i \in \mathcal{I}_r} \mathbf{1}[\hat{y}_i^{\pi} = y_i] \frac{d_i^{\star}(r)}{\max(L_i, d_i^{\star}(r))}, \qquad r \in \{1, 3, 5\}\,\text{m}. \tag{8}$$

Here $\mathcal{I}_r$ is the set of evaluated episodes whose r-conditioned target region is nonempty and reachable, and $N_r = |\mathcal{I}_r|$. The eligible set and denominator are computed separately for each evaluation cohort. The quantity $d_i^{\star}(r)$ is the shortest collision-free path from the initial position to a reachable NavMesh point within 0.1 m of reference distance r, and L_i is the executed path.

Visibility and full-body framing statistics are reported only when the target identity can be estimated without using prohibited ground-truth information.

D TRAINING AND IMPLEMENTATION DETAILS

This section specifies the learned navigator's architecture and behavior-cloning procedure. EAOG and recognizer details follow in Appendix E.

D.1 NAVIGATION ARCHITECTURE

The learned observation policy in Section 5 adapts Uni-NaVid (Zhang et al., 2025) with a five-action movement head. The actions are FORWARD, BACKWARD, LEFT, RIGHT, and WAIT; EAOG makes the separate observation-termination decision. At deployment, the navigator uses RGB, the language instruction, and its own executed actions. Forward and backward request approximately 0.25 m of translation, left and right request a 15° rotation, and wait holds the pose for one 0.25-s control step. The fixed instruction is: *Find the person and maintain a clear, well-framed body view at a comfortable distance. Keep observing useful body motion; front or side views are preferred when reachable without losing the person. Move back when the person approaches, and go around obstacles. Use WAIT when the current body view is stable and useful.* The student does not receive a separate text encoder; the instruction is processed by the Uni-NaVid language-model tokenizer and embedding layer.

Short-window visual and language encoding. Let n index navigation decisions at 4 Hz. At each decision, the navigator uniformly samples up to $F_{\text{nav}} = 6$ frames from the preceding 2 s and resizes them to 224×224. The frozen EVA-CLIP tower produces a class token and a 16×16 grid of 1408-dimensional patch tokens per frame. After removing the class token, each frame is average-pooled to a 2×2 grid for coarse temporal context. The current frame is additionally pooled to an 8×8 grid for finer spatial detail. A full six-frame window therefore supplies 88 EVA tokens ($6 \times 4 + 8 \times 8$).

A multimodal projector maps the visual tokens from the EVA encoder into the language-model embedding space. It consists of two linear layers with dimensions $1408 \to 4096 \to 4096$, separated by a GELU activation. The navigation instruction is formatted using the Vicuna conversation template, tokenized by the language model's tokenizer, and converted into text embeddings by its input embedding layer. The projected visual tokens replace the image placeholder in the token sequence, after which the visual and text embeddings are jointly processed by the LLaMA backbone. The hidden state of the final valid multimodal token forms the fused representation $\mathbf{h}_n^{\text{VLA}} \in \mathbb{R}^{4096}$.

Parameter-efficient adapters. A residual adapter is inserted into each of the last four LLaMA blocks. For the hidden state $\mathbf{H}^{(l)}$ of block l, it computes

$$\widetilde{\mathbf{H}}^{(l)} = \mathbf{H}^{(l)} + \mathrm{Dropout}\Big(\mathbf{W}_{\mathrm{up}}^{(l)}\,\mathrm{GELU}\Big(\mathbf{W}_{\mathrm{down}}^{(l)}\,\mathrm{LN}(\mathbf{H}^{(l)})\Big)\Big), \tag{9}$$

where the down and up projections have dimensions 64×4096 and 4096×64, respectively. Zero initialization of the up projection preserves the pretrained representation when the adapter is first inserted.

Action history and movement prediction. The policy conditions on the two most recently executed actions and the run length r_n of the current repeated action. A shared embedding table with six entries and width 32 represents START and the five actions. A second table with five entries and width 16 represents the run-length bins $\{0, 1, 2\text{–}3, 4\text{–}7, 8+\}$. Their concatenation is

$$\mathbf{x}_n = [\mathbf{h}_n^{\mathrm{VLA}}; \mathbf{e}_a(a_{n-1}); \mathbf{e}_a(a_{n-2}); \mathbf{e}_r(r_n)] \in \mathbb{R}^{4176}. \tag{10}$$

The movement head maps this state to five action logits:

$$\mathbf{z}_n^{\mathrm{nav}} = \mathbf{W}_2\,\mathrm{Dropout}(\mathrm{GELU}(\mathbf{W}_1\,\mathrm{LN}(\mathbf{x}_n))), \tag{11}$$

with layer widths $4176 \to 1024 \to 5$. Greedy inference executes $\arg\max_a z_{n,a}^{\mathrm{nav}}$. Only actions acknowledged by the simulator enter the executed-action history.

In the final training pass, the EVA tower and original 7B language model weights are frozen. Trainable components are the four adapters, action history embeddings, movement head, and final RMSNorm. After training, these components are also frozen for EAOG training and evaluation.

D.2 Supervision Signal and Behavior Cloning

Privileged observation expert. During supervision generation, the expert accesses target world position and motion trend, scene NavMesh, executable paths, person–robot distance, and rendered-body visibility. It balances navigability, standoff distance, full-body visibility, and stable observation. Its actions include backing away from an approaching person, turning to reacquire a lost target, avoiding non-executable translations, and waiting to preserve a useful view. These privileged variables are excluded from student inputs. The expert also checks whether a proposed translation is executable and can leave a state unlabeled when its action supervision is unreliable. The simulator restores the episode's scene instance, matching NavMesh, robot and person initialization, and motion phase before rollout; replayed human motion continues cyclically. At evaluation, none of the expert's geometry, target pose, body visibility, future motion, or emotion label is supplied to the student.

Initial behavior cloning. The expert corpus covers training episodes. An event-balanced view retains 480,000 labeled state–action pairs, selecting 12–40 expert decisions per episode while maintaining coverage of this corpus. Each pair contains the student-observable state and its expert action label; these are decision samples, not additional dataset episodes. Selection preserves initial states, action transitions, backward motion, person approach, visibility changes, target loss and reacquisition, informative-view neighborhoods, and motion-loop boundaries. The final adaptation uses two epochs with effective batch size eight; the frozen visual tower, multimodal projector, and original language-model weights are not updated.

E EAOG Architecture and Training Details

During EAOG training, the navigator, EVA visual tower, and affect recognizer remain frozen. The quality and emotion heads supervise the evidence states, and the termination head selects the window submitted for recognition.

E.1 Candidate Windows and Observation Budget

The navigator acts at $4\,\mathrm{Hz}$. EAOG first evaluates at $12\,\mathrm{s}$ and then once per second through the $48\,\mathrm{s}$ deadline. Each window $\mathcal{W}_t$ contains $F = 32$ deterministically sampled frames from the most recent 48 RGB observations, corresponding to $12\,\mathrm{s}$ of observation. The affect recognizer receives only the selected window, not the recurrent state or the complete observation history.

E.2 Visual Tokens and Emotion Evidence

Visual token construction. EAOG uses the EVA-CLIP visual tower of Uni-NaVid, bypassing the multimodal projector and language model. Removing the class token leaves a 16×16 grid of 1408-dimensional patch features for each frame. Average pooling reduces this to a 2×2 spatial grid. The 32 frames are divided into six contiguous, nearly equal-sized temporal bins. For bin $\mathcal{B}_u$ and spatial cell v, the pooled feature is

$$\mathbf{v}_{t,u,v} = \frac{1}{|\mathcal{B}_u|} \sum_{f \in \mathcal{B}_u} \widetilde{\mathbf{P}}_{t,f,v}, \qquad u = 1, \ldots, 6, \quad v = 1, \ldots, 4, \tag{12}$$

where $\widetilde{\mathbf{P}}$ denotes the spatially pooled EVA features. The resulting 24 tokens retain coarse temporal and spatial structure; averaging does not preserve frame order within a temporal bin.

Token encoding and emotion queries. Each token is normalized, projected to $d = 128$, and given learned temporal and spatial embeddings, $\mathbf{a}_u$ and $\mathbf{b}_v$:

$$\mathbf{x}_{t,u,v} = \mathbf{W}_p \, \mathrm{LN}(\mathbf{v}_{t,u,v}) + \mathbf{a}_u + \mathbf{b}_v. \tag{13}$$

A single-layer, four-head Transformer encoder maps these tokens to $\mathbf{H}_t \in \mathbb{R}^{24 \times d}$. Five learned query vectors form $\mathbf{Q} \in \mathbb{R}^{5 \times d}$. Their fixed output indices correspond to angry, fearful, happy, neutral, and sad through the auxiliary emotion loss. They extract evidence as

$$\mathbf{Z}_t^{(1)} = \mathrm{LN}\big(\mathbf{Q} + \mathrm{MHA}(\mathbf{Q}, \mathbf{H}_t, \mathbf{H}_t)\big), \tag{14}$$

$$\mathbf{Z}_t = \mathrm{LN}\big(\mathbf{Z}_t^{(1)} + \mathrm{FFN}(\mathbf{Z}_t^{(1)})\big). \tag{15}$$

Each row $\mathbf{z}_{t,c}$ is a class-conditioned feature vector used by the recurrent memory and auxiliary emotion readout. The query vectors are randomly initialized trainable parameters. At every candidate time, the same five vectors query that window's encoded visual tokens; the resulting $\mathbf{z}_{t,c}$ changes with the current view. They are not updated by the frozen recognizer's predictions at test time. The auxiliary emotion loss associates their fixed indices with the five labels, while cross-attention lets them learn distinct class-conditioned visual summaries. The pooled tokens preserve coarse within-window temporal and spatial structure, but the model does not retain all 32 frame-level token grids.

Recurrent memory and output heads. The five query streams share a GRU with input and hidden dimension 128, but maintain separate recurrent states:

$$\mathbf{e}_{t,c} = \mathrm{GRU}(\mathbf{z}_{t,c}, \mathbf{e}_{t-1,c}), \qquad c = 1, \ldots, 5. \tag{16}$$

Here $t - 1$ denotes the previous one-second candidate, and the recurrent history is confined to the current episode. The concatenation $\overline{\mathbf{e}}_t = [\mathbf{e}_{t,1}; \ldots; \mathbf{e}_{t,5}]$ has dimension 640. Its readouts are

$$\ell_{t,c}^{\mathrm{emo}} = \mathbf{w}_e^\top \mathbf{e}_{t,c} + b_e, \qquad \widehat{r}_t = f_{\mathrm{quality}}(\overline{\mathbf{e}}_t), \qquad h_t = \sigma\big(g_{\mathrm{term}}(\overline{\mathbf{e}}_t)\big). \tag{17}$$

The quality and termination heads are separate multilayer perceptrons. The quality output $\widehat{r}_t$ predicts standardized recognition risk; it is not an input to the termination head. Both auxiliary heads influence the shared representation during training and are unnecessary for the deployment decision. EAOG contains approximately 0.72M trainable parameters. All five recurrent states start at zero at the beginning of each episode. The streams share GRU weights but have separate hidden states, so a new candidate can change the evidence for each emotion differently. The termination head sees their concatenation. Its h_t is a conditional termination probability reflecting whether further observation remains worthwhile, rather than emotion confidence; the quality and emotion heads are used only for training supervision. The complete EAOG module has 716,163 trainable parameters.

E.3 Recognizer Scores and Training Objectives

Candidate-label scoring. The frozen recognizer described in Sec. 6.1 scores each of the five candidate answers under a fixed affect prompt q. It is the visual-only, 32-frame VideoLLaMA2.1-7B-AV checkpoint at epoch 1, global step 7373, obtained by QLoRA adaptation on the three passive training views P-Init, P-Ref, and P-Mon (19,662 training episodes times three views). The checkpoint

is frozen before query-module supervision, validation, and test evaluation. Its five answer candidates are *Angry*, *Fearful*, *Happy*, *Neutral*, and *Sad*. Let $(w_{c,1}, \ldots, w_{c,L_c})$ be the token sequence for answer a_c, including its end-of-sequence token. Its score and normalized probability are

$$s_{t,c} = \sum_{k=1}^{L_c} \log P_\phi(w_{c,k} \mid \mathcal{W}_t, q, w_{c,<k}), \qquad p_{t,c} = \frac{\exp(s_{t,c}/T_{\text{cal}})}{\sum_{c'=1}^{5} \exp(s_{t,c'}/T_{\text{cal}})}. \tag{18}$$

The same sum-log-probability scores are used for offline supervision and the selected test window. In this equation, $T_{\text{cal}} > 0$ calibrates probabilities for the offline risk expert; the direct recognition interface uses $T = 1$ for its raw five-candidate probabilities. Both have the same argmax. Before the final query, neither form of recognizer probability is available to EAOG. Scores sum token log probabilities, including the end-of-sequence token, without length normalization or free-form generation. The probability-risk expert uses a scalar temperature fitted on training candidates ($T_{\text{cal}} = 1.4424$ in the supplied training record); validation and test labels do not refit it. A positive temperature leaves the predicted class unchanged but affects Half-Brier risk and the soft stopping target. The prompt is fixed by the dataset's target-person protocol: a single-person prompt asks for the visible person's dominant emotion, while dyadic prompts specify the yellow-shirted man, green-shirted woman, or either eligible person. No prompt is rewritten using the ground-truth emotion or an intermediate prediction.

Fixed recognizer prompts. The answer labels and end-of-sequence scoring rule above are shared by four prompt variants selected from the episode's target-person metadata:

Single person: *Watch the video and classify the dominant emotional state of the visible person. Choose exactly one label: Angry, Fearful, Happy, Neutral, Sad. Return only the label.*

Yellow man: *Classify the body emotion of the specified target using only the video. The target is the man wearing a yellow top. Choose exactly one label: Angry, Fearful, Happy, Neutral, Sad. Return only the label.*

Green woman: *Classify the body emotion of the specified target using only the video. The target is the woman wearing a green top. Choose exactly one label: Angry, Fearful, Happy, Neutral, Sad. Return only the label.*

Either actor: *Classify the body emotion of the specified target using only the video. Either of the two visible people is an eligible target. Choose exactly one label: Angry, Fearful, Happy, Neutral, Sad. Return only the label.*

Recognition loss and soft stopping targets. For the one-hot episode label $\mathbf{y}$, the Half-Brier score is

$$R_t = \frac{1}{2} \sum_{c=1}^{5} (p_{t,c} - y_c)^2. \tag{19}$$

This is a probability-prediction loss rather than a classification error indicator. In Eq. 5, time is measured in seconds, $\lambda = 0.002$, and $\tau = 0.005$. The expert assigns more mass to windows with smaller recognition loss and waiting cost. It uses every candidate on the complete training rollout, including future windows relative to an intermediate decision. This privileged expert supplies training targets, while each student prediction uses only the visual prefix available at that candidate time. The supplied training record contains 533,085 legal candidate windows. These counts refer to rollouts and candidate windows, respectively, not additional benchmark episodes. Candidate sequences can be shorter than the 48-s maximum; their padding is masked, and recurrent states are updated only over valid candidates.

First-query distribution. For each rollout, let $t_{\text{end}} = \max \mathcal{T}$ be its last legal candidate (48 s for a complete online episode). The hazards define a distribution over the first query, with remaining probability mass assigned to this forced final candidate:

$$\rho_t = \begin{cases} h_t \displaystyle\prod_{j \in \mathcal{T}: j < t} (1 - h_j), & t < t_{\text{end}}, \\ \displaystyle\prod_{j \in \mathcal{T}: j < t_{\text{end}}} (1 - h_j), & t = t_{\text{end}}. \end{cases} \tag{20}$$

We minimize the cross-entropy from the expert distribution,

$$\mathcal{L}_{\text{term}} = -\sum_{t\in\mathcal{T}} d_t \log \rho_t. \tag{21}$$

The terminal hazard is not needed because stopping at the deadline is forced. Equivalently, this objective is a weighted binary cross-entropy over nonterminal hazards, with target $h_t^* = d_t/S_t$ and weight $S_t = \sum_{j\in\mathcal{T}:j\geq t} d_j$. Thus the sequence objective couples the per-time targets through the remaining expert probability mass.

Auxiliary supervision. Training-candidate statistics standardize the risk as $\widetilde{R}_t = (R_t - \mu_R)/\sigma_R$; the supplied training record gives $\mu_R = 0.328612$ and $\sigma_R = 0.167778$. The two auxiliary losses are

$$\mathcal{L}_{\text{quality}} = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \text{SmoothL1}(\widehat{r}_t, \widetilde{R}_t), \tag{22}$$

$$\mathcal{L}_{\text{emotion}} = -\frac{0.5}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\sum_{c=1}^{5} y_c \log \text{softmax}(\boldsymbol{\ell}_t^{\text{emo}})_c. \tag{23}$$

The complete objective is

$$\mathcal{L}_{\text{EAOG}} = \mathcal{L}_{\text{term}} + \mathcal{L}_{\text{quality}} + \mathcal{L}_{\text{emotion}}. \tag{24}$$

Risk regression encourages features predictive of the recognizer's loss on the current window. Emotion classification associates the five evidence streams with the episode label. The quality target is the risk of the *current* candidate, not the best risk observed later. SmoothL1 regression teaches the shared representation to distinguish usable from weak evidence. The emotion loss gives the five randomly initialized query slots their fixed class indexing; the shared readout is discarded for final recognition. Neither auxiliary head is an explicit input to the termination head. The stopping loss is the only objective that directly trains the first-query distribution. EAOG is optimized with AdamW for 50 epochs using batches of eight episode sequences and gradient clipping at 1.0; candidate padding is masked. The supplied run uses seeds 20260918 and selects the stated training epoch without using test outcomes.

E.4 Online Execution and Query Accounting

During the first 12 s the frozen navigator acquires observations. At each subsequent candidate time, EAOG constructs the current window, updates its recurrent states, and evaluates h_t. If $h_t \geq \theta$, observation ends; otherwise navigation continues until the next candidate. At 48 s the external controller ends observation regardless of the predicted hazard.

E.5 Validation Selection of the Query Threshold

For each trained EAOG variant, we sweep the stopping threshold on validation trajectories and select the value with the highest validation Macro-F1 among those whose mean query time is at most 36 s, or 75% of the 48-s observation horizon. The forced 48-s queries are included when computing this mean. The selected threshold is then fixed for every test episode of that variant; test predictions are not used for threshold selection. The resulting values are $\theta = 0.30$ for EAOG.

The full EAOG test result of 45.47% Macro-F1 at a 33.72-s mean query time in Table E therefore uses $\theta = 0.30$. Because each ablation has its own validation-selected threshold, its test recognition score and mean query time should be interpreted together.

The reported recognition call scores all five candidate answers for the selected window. Online computation also includes the navigation model, EVA encoding, and EAOG updates.

Table D: Macro-F1 (%) on the complete 3,941-episode test split. Within each row, all observation conditions use the same recognizer and episode set. P-Init is the embodied initial observation, P-Ref is the privileged reference observation, and P-Mon is the fixed room-monitor observation. Heur. denotes the training-free heuristic baseline, with paired P-Init fallback when no valid new observation is acquired. Δ_{gain} is the displayed Heur. value minus P-Init.

Perception model	Parameters	P-Mon Monitor view	P-Ref Reference view	P-Init Initial view	Heur. Heuristic baseline	$\Delta_{\text{gain}} \uparrow$ Heur. − P-Init
Open-source general-purpose models						
InternVL2.5 (Chen et al., 2025b)	4B	9.92	20.60	10.04	**12.81**	+2.77
InternVL2.5	8B	11.36	21.90	10.88	**13.51**	+2.63
InternVL2.5 INT8	38B	11.39	19.97	10.74	**13.22**	+2.48
InternVL2.5	78B	20.70	30.02	17.24	**18.52**	+1.28
Qwen2.5-VL (Bai et al., 2025b)	7B	5.99	6.12	6.09	**6.11**	+0.02
Qwen2.5-VL INT8	32B	5.92	6.21	6.12	6.06	−0.06
Qwen2.5-VL BF16	72B	6.63	7.57	6.53	**7.63**	+1.10
Qwen2.5-Omni visual-only (Xu et al., 2025)	7B	7.78	12.09	7.78	**9.45**	+1.67
Qwen3-VL-Instruct (Bai et al., 2025a)	8B	11.79	15.54	10.12	**13.19**	+3.07
Qwen3-VL-Thinking	8B	8.68	11.15	7.93	**9.59**	+1.66
VideoLLaMA2 (Cheng et al., 2024b)	7B	7.47	12.69	7.08	**10.23**	+3.15
VideoLLaMA2	72B	10.25	19.05	9.80	**14.08**	+4.28
VideoLLaMA2.1-16F	7B	7.97	16.25	8.73	**11.20**	+2.47
VideoLLaMA2.1-AV visual-only	7B	20.27	27.65	18.67	**20.89**	+2.22
InternVideo2-Chat (Wang et al., 2024)	8B	7.04	9.98	7.04	**8.07**	+1.03
LongVA-DPO (Zhang et al., 2024)	7B	8.95	13.57	7.55	**10.80**	+3.25
MiniCPM-V2.6 (Yao et al., 2024)	8B	6.34	9.89	6.55	**6.79**	+0.24
Open-source emotion-recognition models						
AffectGPT-frame (Lian et al., 2025)	7B	11.63	10.11	11.79	11.13	−0.66
Emotion-LLaMA (Cheng et al., 2024a)	7B	15.89	25.29	17.16	**19.04**	+1.88
Open-source mean (19)	–	10.31	15.56	9.89	**11.70**	+1.81
Closed-source general-purpose models						
Gemini-3.0-Flash	–	25.11	28.33	24.77	24.62	−0.15
Gemini-3.0-Pro	–	25.21	30.21	22.66	**25.42**	+2.76
GPT-5.2	–	21.08	28.20	20.89	**21.00**	+0.11
GPT-5.6-sol	–	21.53	31.48	19.29	**21.87**	+2.58
Claude Opus 5	–	29.27	35.68	25.46	**28.37**	+2.91
Closed-source mean (5)	–	24.44	30.78	22.61	**24.26**	+1.65

F Additional Experimental Results

F.1 Inference and Execution Settings

Perception models and inference The zero-shot study evaluates 19 open-source and five closed-source model configurations on the test split. A configuration denotes a specific model, scale, quantization setting, and inference interface. These recognizers are used without AffectSim adaptation. The learned-system study uses a separately adapted and frozen visual-only VideoLLaMA2.1-7B-AV recognizer.

Most open-source general-purpose vision-language models use eight uniformly sampled frames resized to 224×224. VideoLLaMA2.1-16F follows the same frame budget, whereas Emotion-LLaMA uses its native single-frame visual input. Closed-source models receive either the full source-duration video or eight uniformly sampled frames according to their supported interface. For each model configuration, the inference interface remains unchanged across all observation conditions.

Zero-shot heuristic baseline The training-free heuristic baseline first uses a frozen Uni-NaVid navigation planner (Zhang et al., 2025) to search for the target person. After handoff, a rule-based controller uses person location and depth to keep the target visible and adjust the observation viewpoint. The baseline uses neither emotion supervision nor recognizer feedback, and has no access to oracle trajectories or simulator masks. In the 24-recognizer zero-shot evaluation, if no valid new observation is acquired, the recognizer receives the paired P-Init observation. Robot actions run at $4\,\text{Hz}$ with $0.25\,\text{m}$ translations and 15° rotations, for at most 192 primitive actions. After handoff, the acquired recognition clip spans twice the source motion duration, matching P-Init and P-Ref.

Table E: Recognition and mean observation-termination time for P-Init, the heuristic baseline, and EAOG using the AffectSim-adapted, frozen VideoLLaMA2.1-7B-AV recognizer. Each EAOG variant uses a validation-selected threshold (Appendix E.5); the full EAOG retains all losses. A dash denotes an unreported metric.

Method / operating point	F1 (%) ↑	Acc. (%) ↑	Half-Brier ↓	Time (s) ↓
Fixed initial view				
P-Init	32.74	33.95	—	—
Heuristic active observation				
Heuristic baseline	34.52	38.24	0.3874	—
Learned active observation with EAOG				
Without emotion loss	42.65	43.96	0.3789	32.44
Without quality loss	44.65	46.04	0.3712	35.28
Without waiting cost	42.10	43.33	0.3842	30.70
EAOG	45.47	46.25	0.3789	33.72

EAOG The frozen learned observation policy runs at 4 Hz. At a legal query time, the adapted recognizer receives 32 frames sampled from the preceding 12 s. Observation ends after one query or at the 48-s deadline. Architecture, expert construction, and optimization details are in Appendices D and E.

F.2 Full Zero-Shot Observation Results

Table D gives the 24 configuration-level scores summarized in main-paper Table 2. The group means average configurations equally.

Observation-quality gap On average, P-Ref yields substantially higher Macro-F1 than P-Init. Across the 19 open-source configurations, mean Macro-F1 rises from 9.89% under P-Init to 15.56% under P-Ref. For the five closed-source models, it rises from 22.61% to 30.78%. P-Ref achieves the highest score for 18 of 19 open-source configurations and 23 of 24 configurations overall.

External monitoring P-Mon provides a fixed room-level external view as an additional diagnostic. For open-source models, P-Mon reaches 10.31% Macro-F1, only slightly above the 9.89% obtained by P-Init and substantially below the 15.56% of P-Ref. Thus, simply providing an external scene-level camera does not reproduce the benefit of a deliberately informative observation trajectory. Observation quality depends jointly on target visibility, framing, orientation, distance, and the discriminative motion cues exposed to the recognizer.

F.3 New-Observation Subset

Heuristic baseline acquired a valid new post-handoff clip in 1,881 of 3,941 test episodes. The complete-protocol result also includes 2,060 episodes with P-Init fallback.

Observation gap recovery Using Observation Gap Recovery (OGR) computed before score rounding, heuristic baseline recovers 32.0% of the mean P-Ref–P-Init gap for open-source models and 20.1% for closed-source models. Active movement therefore closes a measurable, although still incomplete, portion of the gap to the privileged reference observation. On the new-observation subset, heuristic baseline outperforms P-Init for 20 of 24 configurations (main-paper Figure 4).

F.4 Stopping-Time Controls and EAOG Comparison

The first legal fixed query time is 12 s; the fixed-time controls displayed in main-paper Table 4 query at 24, 30, or 48 s. Uniform random samples one of the 37 integer candidate times from 12 through 48 s. All use the same frozen learned observation policy and AffectSim-adapted, frozen

Table F: Macro-F1 (%) on the complete test set and the 1,881-episode new-observation subset defined by successful post-handoff acquisition. In the remaining 2,060 episodes, heuristic baseline uses the paired P-Init input. Gain denotes Heur.−P-Init.

Model group	Evaluation subset	Episodes	P-Init	Heur. Zero-shot policy	Heur.−P-Init
Open-source (19)	Complete test set	3,941	9.89	**11.70**	**+1.81**
	New-observation subset	1,881	10.95	**14.35**	**+3.40**
Closed-source (5)	Complete test set	3,941	22.61	**24.26**	**+1.65**
	New-observation subset	1,881	25.38	**28.29**	**+2.91**

Table G: Recovery Rate and E-SPL@r (%) on the complete test split. Recovery Rate measures the fraction of incorrect P-Init predictions corrected under heuristic baseline. E-SPL weights final heuristic baseline recognition correctness by path efficiency relative to a distance-conditioned target region at $r \in \{1, 3, 5\}$ m. Empty or unreachable target regions are excluded, yielding 3,811, 3,887, and 3,859 valid episodes at 1, 3, and 5 m, respectively.

Perception model	Parameters	Recovery Rate	E-SPL @1 m	E-SPL @3 m	E-SPL @5 m
Open-source general-purpose models					
InternVL2.5	4B	3.24	1.33	0.68	0.70
InternVL2.5	8B	2.92	1.29	0.67	0.60
InternVL2.5 INT8	38B	2.98	1.23	0.62	0.63
InternVL2.5	78B	4.05	1.74	1.01	0.67
Qwen2.5-VL	7B	0.24	0.13	0.08	0.09
Qwen2.5-VL INT8	32B	0.15	0.05	0.02	0.03
Qwen2.5-VL BF16	72B	1.16	0.58	0.29	0.31
Qwen2.5-Omni visual-only	7B	1.73	0.71	0.37	0.34
Qwen3-VL-Instruct	8B	5.74	2.68	1.49	1.28
Qwen3-VL-Thinking	8B	2.22	1.11	0.65	0.49
VideoLLaMA2	7B	2.77	1.24	0.65	0.71
VideoLLaMA2	72B	4.68	2.00	0.99	0.95
VideoLLaMA2.1-16F	7B	2.59	1.13	0.53	0.62
VideoLLaMA2.1-AV visual-only	7B	7.76	3.43	1.90	1.54
InternVideo2-Chat	8B	1.17	0.51	0.25	0.31
LongVA-DPO	7B	3.50	1.65	0.99	0.89
MiniCPM-V2.6	8B	0.79	0.42	0.29	0.24
Open-source emotion-recognition models					
AffectGPT-frame	7B	3.31	1.32	0.63	0.74
Emotion-LLaMA	7B	7.88	3.03	1.75	1.43
Open-source mean (19)	–	3.10	1.35	0.73	0.66
Closed-source general-purpose models					
Gemini-3.0-Flash	–	6.08	2.37	1.39	0.97
Gemini-3.0-Pro	–	9.11	3.93	2.34	1.52
GPT-5.2	–	4.23	1.75	1.05	0.78
GPT-5.6-sol	–	5.51	2.27	1.41	0.86
Claude Opus 5	–	11.27	4.83	2.66	2.03
Closed-source mean (5)	–	7.24	3.03	1.77	1.23

VideoLLaMA2.1-7B-AV recognizer as EAOG, with 32-frame input on the same learned-system test split. The 24-s rule bisects the full 48-s budget; 30 s bisects the legal candidate interval. Each method takes one recognition query from the preceding 12 s of egocentric video and stops observation at its chosen time.

Main-paper Table 4 reports Macro-F1, OGR, and E-SPL@3 for the displayed timing controls and EAOG. E-SPL@3 follows Eq. 8; its eligible set is defined for the learned-system evaluation cohort

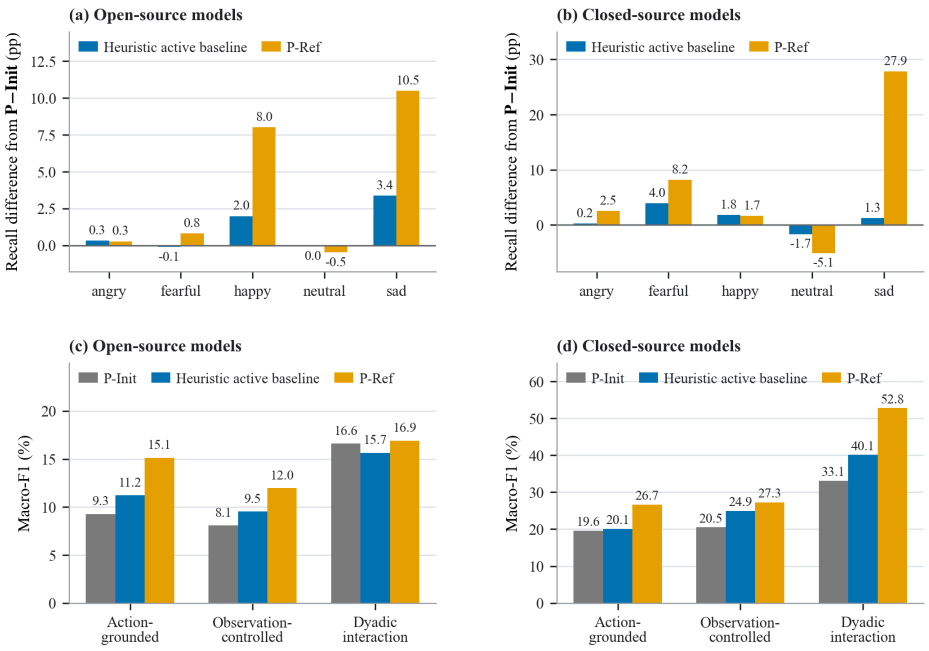


Figure B: Class- and episode-family-level diagnostics on the complete 3,941-episode test set. (a,b) Mean per-class recall gains of heuristic baseline and P-Ref relative to P-Init across the 19 open-source and five closed-source model configurations. (c,d) Mean Macro-F1 for action-grounded single-person ($n = 2{,}899$), observation-controlled single-person ($n = 832$), and dyadic-interaction ($n = 210$) episodes. Each value is an unweighted mean of model-wise scores.

Table H: Execution diagnostics of the heuristic baseline on the complete test split. Mean path is computed per episode; handoff and budget exhaustion are percentages of episodes; pre-handoff path share is pooled over total locomotion; and clipped moves are normalized by attempted translations.

Episodes	Stable handoff (% episodes)	Mean path (m/episode)	Path before handoff (% locomotion)	Budget exhaustion (% episodes)	Clipped moves (% attempts)
3,941	47.73	23.23	97.4	52.27	14.10

separately from the zero-shot eligible counts in Table G. EAOG has the highest Macro-F1 and E-SPL@3 among the reported operating points.

Table E reports P-Init and heuristic-baseline references, as well as accuracy, Half-Brier score, and mean termination time for the EAOG variants in main-paper Table 5. The heuristic row uses rule-based viewpoint control after handoff and has no reported termination-time value. In this adapted-recognizer comparison, a failed handoff submits the available egocentric window; the zero-shot heuristic evaluation in Table D uses paired P-Init fallback.

F.5 Class and Episode-Family Diagnostics

The lower panels compare P-Init, heuristic baseline, and P-Ref across action-grounded, observation-controlled, and dyadic-interaction episodes, using unweighted means of model-wise Macro-F1. Heuristic baseline improves the open-source action-grounded and observation-controlled means from 9.3% to 11.2% and from 8.1% to 9.5%, respectively, but lowers the dyadic mean from 16.6% to 15.7%. Closed-source heuristic baseline improves all three families, most strongly for dyadic interaction (33.1% to 40.1%); P-Ref reaches 52.8% there. Thus, the aggregate gains do not imply uniform benefit across episode families.

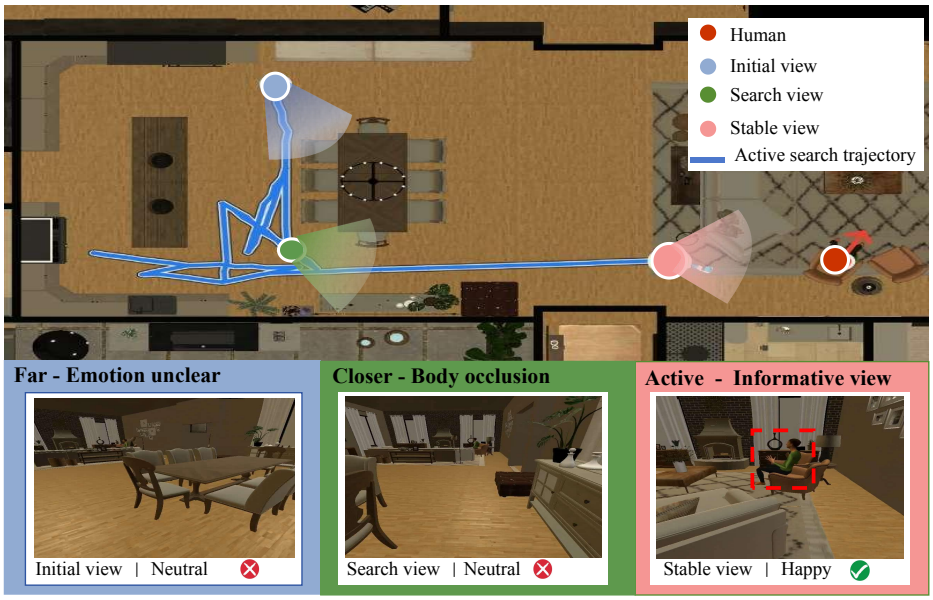


Figure C: Example of active observation for emotion recognition. The top-down map shows the performer, sampled robot viewpoints, and executed trajectory. The distant initial observation is incorrectly recognized as neutral; moving closer still leaves the body partially occluded, while further repositioning provides a less-occluded observation that is correctly recognized as happy.

F.6 Recovery, Efficiency, and Qualitative Example

Table G gives Recovery Rate and E-SPL for all 24 zero-shot recognizer configurations and their open- and closed-source group means. E-SPL evaluates recognition together with path efficiency relative to target regions at 1, 3, and 5 m. These values use heuristic baseline and its eligible-episode counts, separately from the learned navigator evaluated in main-paper Table 4.

Episode-level recovery Claude Opus 5 obtains the highest Recovery Rate at 11.27%, while Emotion-LLaMA achieves the highest value among the open-source models at 7.88%. These results provide an episode-level view complementary to the aggregate Macro-F1 gains.

Path-aware recognition E-SPL evaluates correct heuristic baseline predictions together with the path efficiency of observation acquisition. Across the 24 model configurations, mean E-SPL is 1.70%, 0.95%, and 0.78% at target distances of 1, 3, and 5 m, respectively. Claude Opus 5 obtains the highest E-SPL at all three distances, while VideoLLaMA2.1-AV achieves the highest values among the open-source models.

Table H reports execution diagnostics for heuristic baseline, whose acquisition budget and fallback protocol differ from the learned-system evaluation.

Across all 3,941 test episodes, 47.73% reach stable handoff and mean path length is 23.23 m. Pre-handoff movement accounts for 97.4% of locomotion, 52.27% of episodes exhaust the original budget, and 14.10% of translation attempts are clipped by the NavMesh.

Figure C shows an episode in which repositioning reveals affective body-motion cues that approaching alone does not expose.

G Limitations, Ethics, and Future Work

G.1 Limitations

AffectSim has several limitations. First, the current benchmark relies primarily on acted rather than spontaneous emotional expressions. Although acted performances provide reproducible affective behaviors, they may not fully capture the diversity and subtlety of naturally occurring emotion. Second, avatar rendering and motion retargeting do not establish real-world transfer; the current evaluation supports controlled within-benchmark conclusions about observation and policy behavior. Third, the label space contains five discrete emotion categories and does not represent mixed, continuous, or culturally dependent affective states. Fourth, simulated humans follow predefined motions and do not react to the observing agent. AffectSim therefore focuses on embodied affective perception under controllable observation rather than fully closed-loop affective interaction or deployment performance.

G.2 Ethics and Broader Impact

Emotion perception technologies raise important ethical concerns. The motion capture protocol follows procedures for informed consent, compensation, data governance, withdrawal, and data use. Released avatars do not reproduce the participants' real facial appearance, reducing direct identity exposure compared with releasing in-the-wild human video.

Nevertheless, emotion inference may be misused for surveillance, manipulation, or psychological profiling. The benchmark is intended for research on embodied perception, assistive robotics, and human-centered AI rather than high-stakes decision making or surveillance. Release documentation will describe the subjectivity and cultural limitations of emotion labels, intended uses, and procedures for reporting potential misuse.

G.3 Future Work

Future versions may incorporate spontaneous affective behavior, finer-grained or continuous affective states, multimodal cues, person-specific expression styles, more diverse avatars and environments, and richer social configurations. Reactive virtual humans would further allow the person's behavior to change in response to the observer. Real-robot studies can then test whether policies learned from the controlled benchmark retain their observation strategies under real sensing and dynamics, treating transfer as a separate empirical question rather than an assumption of the benchmark.