Proceedings of CONF-CDS 2026 Symposium: Machine Learning and Neural Network Applications in Engineering
DOI: 10.54254/2755-2721/2026.AST36699

From System Compensation to Semantic Adaptation: A Critical Survey of Latency Mitigation in Vision-Based Gesture Interaction

Ziyuan Wu

Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Foshan, China
2513332367@qq.com

Abstract. Vision-based gesture pipelines run sub-30 ms per frame, yet closed-loop interaction often exceeds ~200 ms. This study treats latency as an end-to-end property spanning perception, inference, decision and feedback. This critical survey synthesizes a purposively screened corpus of gesture and adjacent interactive systems (2020–2026), plus foundational anchors. This study uses a three-layer taxonomy: system compensation, semantic adaptive inference, and hybrid semantic–system coordination. This study finds that system-level remedies dominate MediaPipe-era pipelines but remain blind to recognition uncertainty. Semantic adaptivity optimizes compute under uncertainty but is rarely bound to interactive latency service level objectives (SLOs). The missing link is the runtime interface between confidence and load. This paper proposes a three-layer taxonomy that defines hybrid coordination as a missing layer, identifies the perception–E2E latency gap as the central empirical blind spot, and formalizes Load–Confidence Coupling as a testable proposition with falsification criteria.

Keywords: gesture interaction, latency mitigation, semantic adaptive inference, human-computer interaction, confidence-load coupling

1. Introduction

1.1. User-experience bottlenecks motivating the survey

Vision-based gesture games and motion-sensing interfaces suffer response lag and recognition errors under load [1]. MediaPipe-style pipelines report high frame rates under ideal conditions [2, 3], yet closed-loop delays reach hundreds of milliseconds [4], though per-frame processing is tens of milliseconds [5]. Input-to-display lag degrades pointing [6] and steering [7], while uncertain recognition relies on fixed thresholds rather than confidence- and load-dependent behavior [2].

DOI: 10.54254/2755-2721/2026.AST36699

1.2. Inclusion criteria, keyword strategy and PRISMA-style screening

This work is a critical, taxonomy-driven synthesis, not an exhaustive catalog. Candidates came from a curated working set, targeted queries (ACM, IEEE Xplore, and Google Scholar, 2020-2026), and backward snowballing [1, 8]. Screening of 112 records finalized 23 gesture-interaction papers plus 8 cross-domain anchors (Figure 1).

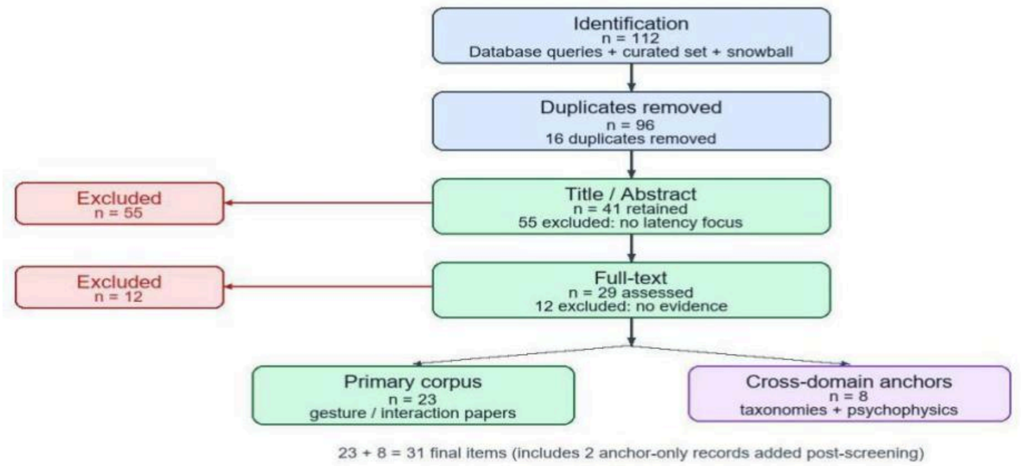

Figure 1. PRISMA-style literature screening flowchart

1.3. Core review questions

RQ1: How do existing studies reduce or mask gesture-interaction latency?

RQ2: Why do system-level compensation and semantic adaptive inference remain weakly connected?

RQ3: Which shared gaps (benchmarks, confidence–load coupling) recur across the corpus?

2. Background and a three-layer taxonomy of latency mitigation

Figure 2 shows the three-layer taxonomy, and Figure 3 shows how mitigation traditions assemble into gesture pipelines.

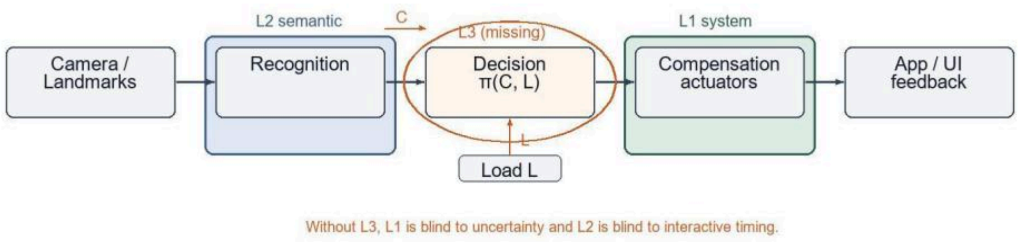

Figure 2. Three-layer taxonomy as a runtime architecture. L3 (dashed) marks the missing confluence of confidence C and load L at decision time

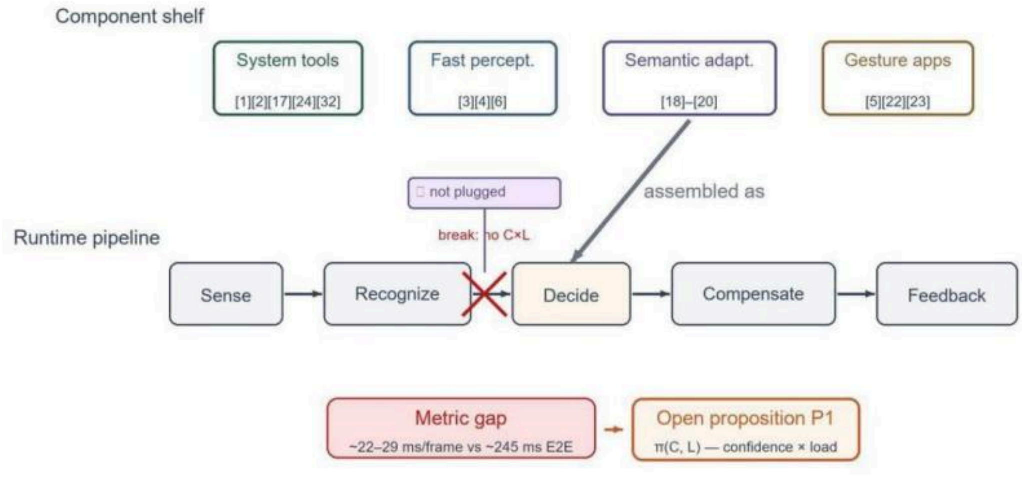

Figure 3. How latency-mitigation traditions assemble into gesture pipelines, where the C×L interface breaks, and the open proposition P1

2.1. System-level compensation: prediction, degradation, filtering and time manipulation

System-level compensation originates in networked game engineering [9]: software that mitigates observable delay by manipulating inputs, predicted states, feedback, or world parameters. Core motifs include Dead Reckoning, input prediction, time manipulation, and filtering [1]; the One Euro filter's jitter–lag trade-off [10, 11] is a concrete remedy.

2.2. Semantic-level adaptive inference: early exiting, confidence skipping and reflective control

Semantic-level adaptive inference allocates computation by uncertainty: early-exit networks terminate when confident enough [12, 13], and Self-RAG gates extra evidence [14]. In gesture, semantic adaptivity is scarce, limited to efficiency neighbors (CNN-LSTM [15], MediaPipe pruning [2], MDRA [16]).

2.3. Hybrid semantic–system coordination: why a third layer is needed

Hybrid coordination would map semantic uncertainty into runtime choices—stronger prediction when confidence is low, lighter models when load spikes, deferred confirmation when both are high. The literature offers only weak neighbors [11], none closing the loop.

2.4. Comparison dimensions used in this survey

Comparison dimensions are listed in Table 1.

DOI: 10.54254/2755-2721/2026.AST36699

Table 1. Cross-domain transferability summary: what moves as a metaphor, what breaks, and what gesture interaction still needs [17]

Source domain	Mechanism	Transfers?	Boundary / why it breaks	Implication for gesture
NLP early exiting	Terminate inference when confident	Metaphor √	Discrete token confidence ≠ continuous motion uncertainty	Use confidence gating as a metaphor
Adaptive NN	Select network depth per example	Metaphor √	No gesture-tailored latency SLO, wrong exit injects false triggers	Needs latency-aware exit criteria
Self-RAG	Reflective retrieval/critique gating	Metaphor √	Optional retrieval violates closed-loop steering	Reflection is too slow for direct manipulation
Dead Reckoning	Predict remote state from timing estimate	Partial √	Known transport delay ≠ intermittent tracking loss	Short-horizon local extrapolation only
One Euro filter	Speed-adaptive jitter–lag trade-off	Direct √	Adapts by motion speed, not confidence or load	Plug into landmark stream, not C×L coupling

3. Confidence-to-decision mapping across studies

Most MediaPipe-era gesture games encode binary finger states—Dino maps a fist to the spacebar [18], Gesto-Drive uses pinch-distance thresholds [5], GestureFlow maps index orientation to steering [19]. These are cheap but fail under occlusion and cannot express graded uncertainty.

Gesture chess interprets the same pinch differently when selecting versus moving [4]; GestureFlow's accuracy degrades from PC to Raspberry Pi [19], yet systems rarely feed load back into thresholds.

Learned controllers (e.g., RL policies) appear more in adjacent communities; within the corpus, learning serves recognition (CNN–LSTM [15], OpenPose+GRU [20], MDRA [16]) rather than meta-decisions.

Early exiting, reflective gating, Dead Reckoning, and One Euro filtering transfer as design metaphors [10, 11]. But gesture confidence is continuous and autocorrelated, so a wrong early exit can inject false triggers.

DOI: 10.54254/2755-2721/2026.AST36699

4. Dynamic thresholding, perceptual masking and evaluation practice

Truly dynamic thresholds—updated online—are uncommon; the corpus offers cooldown timers [4], dwell-time switches [5], speed-adaptive filtering [11], and sliding-window votes [20].

Perceptual masking improves experience without reducing physical latency: latency concealment plays local effects [1], One Euro filtering trades jitter against lag [10, 11], and VR extends masking to spatial uncertainty [21, 22].

4.1. Standardized comparison of reported timing evidence

Table 2. Cross-paper timing and decision comparison

System / Paper	Perception latency	E2E latency	Hardware	Confidence metric	Compensation / decision	Eval. metrics
Gesto-Drive [5]	√ ~22–29 ms/frame	× not reported	Mid-tier GPU/PC	Geometric (pinch px), no calibrated p(class)	Fixed thresholds + dwell mode switch	Acc. ~89–94%, per-frame ms
Gesture chess [4]	Track ~85 ms, classify ~30 ms	⚠ ~245 ms (logic ~130 ms)	Webcam + desktop	Stage FSM ≠ confidence	Context-by-game-stage, EMA α=0.3	E2E decomposition, demo
MediaPipe embedded [2]	FPS / model-size oriented	× not reported	Embedded / edge	Deployment "adaptive" ≠ L2	Compression / optimization	Acc., FPS, footprint
RGB virtual model [11]	Filter-dependent	× not reported	Single RGB camera	Motion speed (One Euro)	One Euro jitter–lag trade-off	Interaction demo, filter quality
GestureFlow [19]	Platform-dependent	× not reported	PC / Jetson / Raspberry Pi	× not reported confidence	Fixed gesture→control map	Acc./responsiveness drift by platform
Dino gesture [18]	× not reported precise ms	× not reported	Webcam demo	Binary fist/open	Fixed mapping	Functional demo, robustness notes

Note: √ = reported, × = not reported, ⚠ = not comparable. Perception latency ≠ end-to-end (E2E) action latency.

Table 2 provides a standardized cross-paper comparison of reported timing evidence and decision mechanisms.

4.2. Datasets, toolchains and benchmark gaps

Recognition research benefits from public dynamic gesture benchmarks [16]; interaction latency research does not—gesture games use ad-hoc webcam setups under incomparable conditions [18, 19].

4.3. Experimental design commonalities and limits

Reported designs show common limits: missing P99/jitter, absent NASA-TLX [23] or delay questionnaires, small samples in rehab demos [24, 25], and evaluation on scripted clips [16]. CHI-level systems evaluate more densely [22]; engineering papers remain demonstration-heavy [26].

5. Critical discussion: diagnosing why the gap persists

The gap is a missing typed runtime contract between confidence signals and compensation actuators: recognition outputs landmarks, classes, or probabilities, while system modules decide with fixed geometry, stage machines, or filters.

The corpus lacks unified benchmarks and relies on incommensurable metrics: authors optimize frame latency while end-to-end action latency remains unreported, so confidence–load coupling is unmeasurable [2, 19].

Hybrid coordination is sparse because of venue silos, toolchain defaults, safety asymmetry, and missing shared tasks: separate venues, demo-sufficient thresholds [3], and salient false triggers [5].

The primary corpus is modest (~23 focal papers) and venue quality is uneven [2]; this survey claims theoretical saturation, not comprehensive coverage.

6. Future directions

Proposition P1 (Load–Confidence Coupling). Let C_t denote recognition confidence and L_t a load vector, and let π(C_t, L_t) select among {predict, smooth, exit early, confirm, degrade}. Policies π exist that improve a joint objective J = α·FalseTrigger + β·MissedTrigger + γ·PerceivedLag + δ·TailLatency relative to confidence-blind baselines [4, 5, 11].

P2 enables P1: a minimal reporting standard should mandate end-to-end action latency (mean/P95/P99), per-frame latency, confidence metric, and decision logic.

Hand-to-avatar mapping research optimizes control precision, structural similarity, and comfort [27]; elicitation studies show users switch between mimicry and symbolic gestures [28]. Latency mitigation should be judged by whether intent is delivered on time, not only FPS [21, 22].

7. Conclusion

This critical survey asked how existing research mitigates latency in vision-based gesture interaction, and why system-level and semantic-level traditions remain weakly connected. By synthesizing a purposively screened corpus plus foundational anchors, this survey organized the field into system-level compensation, semantic-level adaptive inference, and an emerging hybrid layer of semantic–system coordination.

Comparative evidence suggests that classical system techniques improve responsiveness without modeling recognition uncertainty [19], whereas semantic adaptivity optimizes computation under uncertainty but is seldom specialized for real-time motion sensing. MediaPipe-era systems demonstrate that perception front-ends can be fast [5], yet closed-loop delays and brittle thresholds

DOI: 10.54254/2755-2721/2026.AST36699

persist [17]. Across themes—confidence-to-decision mapping, threshold policies, perceptual masking, and evaluation practice—the most consistent gap is the missing dynamic coupling between semantic confidence and system load, compounded by the absence of shared gesture-latency benchmarks and by incommensurable metrics.

Accordingly, this paper's contribution is taxonomic and critical rather than algorithmic: a cross-domain mapping with transferability bounds, a three-layer classification that supersedes a binary split, a metric checklist and comparison table, and a formalizable proposition P1 on confidence–load coupling. Future work should prioritize testing P1 under shared benchmarks and interface principles, this survey intentionally stops short of proposing or validating any new system.

References

[1] Liu, S., Xu, X., & Claypool, M. (2022). A survey and taxonomy of latency compensation techniques for network computer games. *ACM Computing Surveys, 54*(11s), Article 243.

[2] Nirale, I., Ziaullah, M., & Hosamani, R. (2025). MediaPipe-driven real-time hand gesture recognition: An embedded AI framework for adaptive human-computer synergy. *Proceedings of the IEEE GCWCN*, 1–6.

[3] Lugaresi, C., et al. (2019). MediaPipe: A framework for building perception pipelines. arXiv:1906.08172

[4] Reddy, M. M. K., Uyyala, R., Vurubindi, P., & Charan, R. S. (2025). Low latency gesture control chess with multiplayer support. *Proceedings of the IC-CGU*. https://doi.org/10.1109/IC-CGU67042.2025.11338050

[5] Maru, H., Kothari, R., Sanghrajka, B., Mangle, A., & Shaikh, T. (2025). Gesto-drive – Real-time HCI for peripheral-free gaming. *Proceedings of the ICECER*, Antananarivo, Madagascar. https://doi.org/10.1109/ICECER65523.2025.11401101

[6] MacKenzie, I. S., & Ware, C. (1993). Lag as a determinant of human performance in interactive systems. *Proceedings of the INTERCHI '93*, Amsterdam, Netherlands, 488–493.

[7] Accot, J., & Zhai, S. (1997). Beyond Fitts' law: Models for trajectory-based HCI tasks. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Atlanta, GA, USA, 295–302.

[8] Suzuki, R., Karim, A., Xia, T., Hedayati, H., & Marquardt, N. (2022). Augmented reality and robotics: A survey and taxonomy for AR-enhanced human-robot interaction and robotic interfaces. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–33.

[9] Bernier, Y. (2001). Latency compensating methods in client/server in-game protocol design and optimization. *Game Developers Conference (GDC)*. https://developer.valvesoftware.com/wiki/Latency_Compensating_Methods_in_Client/Server_In-game_Protocol_Design_and_Optimization

[10] Casiez, G., Roussel, N., & Vogel, D. (2012). 1 € filter: A simple speed-based low-pass filter for noisy input in interactive systems. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2527–2530.

[11] Wu, D., Huang, J., Zheng, M., & Li, Y. (2023). Virtual model interaction based on single RGB camera. *Proceedings of the NTCI*. https://doi.org/10.1109/NTCI60157.2023.10403685

[12] Teerapittayanon, S., McDanel, B., & Kung, H. T. (2016). BranchyNet: Fast inference via early exiting from deep neural networks. *Proceedings of the ICPR*, 2464–2469.

[13] Bolukbasi, T., Wang, J., Dekel, O., & Saligrama, V. (2017). Adaptive neural networks for efficient inference. *Proceedings of the 34th International Conference on Machine Learning, PMLR 70*, 527–536.

[14] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *Proceedings of the 12th International Conference on Learning Representations*. https://openreview.net/forum?id=hSyW5go0v8, arXiv:2310.11511

[15] Pushparani, S., Muthukrishnan, A., Poongavanam, N., Kumar, K. S., & Arun, M. (2024). Continuous gesture recognition in human-computer interaction through real-time machine learning models. *Proceedings of the 2024 International Conference on Recent Advances in Science, Engineering and Technology*, 1–5.

[16] Liao, G., Huang, L., Chen, K., Gu, Y., & Zhu, T. (2026). MDRA: A motion-guided dual-stream recurrent attention framework for dynamic hand gesture recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications, 22*(5), 1–19.

[17] Mohamed, A. S., Hassan, N. F., & Jamil, A. S. (2024). Real-time hand gesture recognition: A comprehensive review of techniques, applications, and challenges. *Cybernetics and Information Technologies, 24*(3), 163–181. https://doi.org/10.2478/cait-2024-0031

[18] Dinesh, B., Reddy, C. V. N., & Senthamilarasi, N. (2025). Computer vision-based hand recognition and gesture control for Dino games. *Proceedings of the 2025 International Conference on Emerging Smart Computing and Informatics*, March 05–07.
[19] Bharatula, S. D., Vadhegar, U. R., & Maiti, M. (2024). GestureFlow: A novel hand gesture control system for interactive gaming. *Proceedings of the ICCCNT*, 1–6.
[20] Bagladi, M. Z., Gulyás, L., & Szalay, G. (2025). Fast real-time pipeline for robust arm gesture recognition. *Proceedings of Intelligent Robotics FAIR (IntRob'25)*, Budapest, Hungary, 138–143.
[21] Morimoto, K., Kasahara, N., Yamanaka, S., Miyashita, H., & Watanabe, K. (2025). Enhancing freehand VR interaction using fingertip deformation on user performance. *Proceedings of VRST '25*. Montreal, QC, Canada, https://doi.org/10.1145/3756884.3765975
[22] Hwang, S., Kang, S., Oh, J., Park, J., Shin, S., Luo, Y., DelPreto, J., Lee, S., et al. (2025). TelePulse: Enhancing the teleoperation experience through biomechanical simulation-based electrical muscle stimulation in virtual reality. *Proceedings of CHI*, 1–26.
[23] Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology, 52*, 139–183.
[24] Acharya, V., Roy, H., Kamath, S., & Acharya K., A. (2025). Design and development of a virtual gaming console using deep learning techniques for hand rehabilitation. *Proceedings of ICRAIS*, 213–217.
[25] Sivamurugan, C., Yashaswini, K., Nihith, J., Kumar, K., Nagarjuna, K. V., & Gowsalya, P. (2026). Computer vision and gesture-based biofeedback: A serious gaming approach for hemiparesis muscle coordination. *Proceedings of the ICSFT*, 1–5.
[26] Adilzhan, A., & Omarov, N. (2024). Development of augmented reality game using computer vision technology. *Proceedings of the IEEE SIST*, 386–391.
[27] Jiang, Y., Li, Z., He, M., Lindlbauer, D., & Yan, Y. (2023). HandAvatar: Embodying non-humanoid virtual avatars through hands. *Proceedings of the CHI*, 1–17.
[28] Hung, C. W., Chang, R. C., Chen, H. S., Liang, C. H., Chan, L., & Chen, B. Y. (2022). Puppeteer: Exploring intuitive hand gestures and upper-body postures for manipulating human avatar actions. *Proceedings of the VRST*. https://doi.org/10.1145/3562939.3565609