Understanding Emergent Non-Verbal Communication in the Delta Force Competitive Video Game through Multimodal AI Analysis

Jinyuan Guo
2873678510@qq.com
Duke Kunshan University
Kunshan, China

Josef Spjut
jspjut@nvidia.com
Duke University
Durham, USA
NVIDIA
Durham, USA

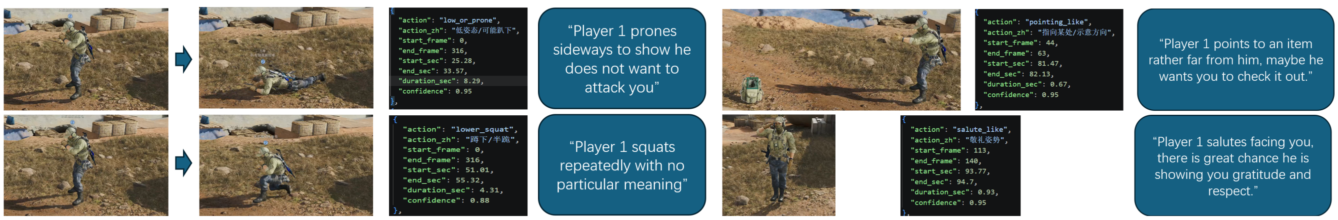

Figure 1: In extraction shooter games like Delta Force, non-verbal communication is central to emergent gameplay, particularly communicating friendly intents. We propose a pipeline using multimodal AI methods on video input to detect different non-verbal gestures as shown in the screenshots. These detected gestures are time aligned with the source video to enable further analysis, such as the example captions on the right.

CCS Concepts

• Applied computing → Computer games; **• Computing methodologies → Computer vision**; **Neural networks**.

Keywords

computer games, nonverbal communication, computer-supported cooperative work

ACM Reference Format:
Jinyuan Guo and Josef Spjut. 2026. Understanding Emergent Non-Verbal Communication in the Delta Force Competitive Video Game through Multimodal AI Analysis. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Technical Workshops (SIGGRAPH Technical Workshops '26), July 19–23, 2026, Los Angeles, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3799828.3816006

1 Introduction

Non-verbal communication plays a critical role in multiplayer games, players often rely on gestures, movement patterns, item interactions, and UI signals to communicate intent, negotiate cooperation willingness, and avoid conflict. May et al. [May et al. 2013] demonstrated how non-verbal game behaviors are effective at changing player behavior and task success.

SIGGRAPH Technical Workshops '26, Los Angeles, CA, USA

Many studies have considered non-verbal communication in competitive esports, though the bulk of them have focused on MOBA games like League of Legends [Leavitt et al. 2016], Dota 2 [Wuertz et al. 2017], and Heroes of the Storm [Zheng and Farzan 2023; Zheng et al. 2023]. MOBA games include a ping feature, that places a mark on the map with intention to communicate warning or helping concepts. While players develop their own intent with these tools, ping largely sticks to the game designer's intended meaning, and are somewhat more limited in scope than more expressive non-verbal communication. Another line of research has also leveraged MOBAs and online matchmaking that matches team mates with strangers to evaluate how ad hoc teams work [Kou and Gui 2014; Lee et al. 2025; Tan et al. 2022].

However, other genres, like extraction shooter and free-for-all battle royale games, have enabled even more ad hoc teaming, betrayals, and rescue behavior. In most shooter games, teams generally coordinate through verbal communication such as map callouts in Counter-Strike [Rusk et al. 2024; Tang et al. 2012]. In contrast to these prior studies, we endeavor to study non-verbal communication in the shooter game genre.

In our work, we analyze gameplay videos to extract non-verbal communication. Some past work has shown that training a model with data across a range of shooter games it is capable of learning some cross-game features due to the similarities within the genre.[Rašajski et al. 2024]. Perhaps the closest related work to our method is GELID, which performs a somewhat different set of operations, including segmenting based on captions, categorizing and grouping with image features and color analysis [Guglielmi et al. 2023]. Others have shown more general video game understanding using VLMs [Taesiri and Bezemer 2025]. We believe our pipeline to be novel, while being composed of relatively well-understood components.

1.1 Delta Force

In the game Delta Force (Warfare), the game follows a search-engage-extract loop similar to games like Escape from Tarkov or ARC Raiders. Since all items are lost upon death, survival and extraction are the primary goals. This creates situations that are not strictly zero-sum. Avoiding conflict can increase both parties' chances of survival, making cross-team coordination beneficial.

Although Delta Force includes voice channels within teams, communication between unaffiliated teams is limited. In this setting, gestures are not simply a replacement for speech, but also a supplementary and sometimes safer layer of signaling. These include behaviors such as signaling surrender through movement, forming temporary alliances, or indicating friendliness by dropping items or using publicly broadcast voice lines (e.g., celebratory phrases or non-threatening signals). Such interactions are not explicitly encoded by the game mechanics but are widely recognized and interpreted by experienced players.

Understanding non-verbal communication patterns is important for two reasons. First, they represent a meaningful layer of player experience that is currently inaccessible to newcomers or external viewers. Second, given their subtlety, gestures provides a testbed for evaluating the capabilities of AI systems in understanding gaming behaviors in complex and dynamic environments.

In this work, we propose a multimodal pipeline for detecting and interpreting emergent non-verbal communication in gameplay videos and results from a preliminary implementation. The pipeline combines pose estimation, visual-textual extraction, object and audio cues, and LLM-based reasoning to transform raw gameplay videos into timestamp-aligned interpretations of player intent. Through Delta Force as a case study, we make three contributions: (1) A taxonomy of gestures used by players to signal friendliness or cooperation. (2) A modular AI pipeline for identifying such gestures from gameplay video. (3) Discussion that this AI pipeline can support empirical studies of social behavior in games.

2 Delta Force Non-verbal Communication

While hostile non-verbal gestures in first-person shooter games often appear as shooting at the first glance, we categorize gestures into two types: atomic gestures and composite gestures. Atomic gestures are single, predefined actions available in the game (emotes or item interactions), while composite gestures emerge from repeated or combined movements performed by players. We contribute the interpretations in Table 1, which were constructed from observation of publicly available Delta Force gameplay videos, and the first author's own gameplay experience with the extraction mode.

2.1 Gesture Recognition Pipeline:

To analyze non-verbal communication in gameplay videos, we designed a multimodal gesture recognition pipeline that extracts and integrates motion, visual, and audio signals into interpretable interaction segments, as illustrated in Figure 2.

The pipeline starts from raw gameplay video. We apply pose estimation (MMpose [Contributors 2020]) to extract skeletal keypoints of avatars over time. Based on these skeletal outputs, we implement a rule-based mapping from low-level motion data to interpretable gestures. Each gesture is defined by a set of consistent spatial and temporal features such as joint angles and repetition patterns, allowing us to map skeletal motion sequences to predefined gesture categories (see Table 1).

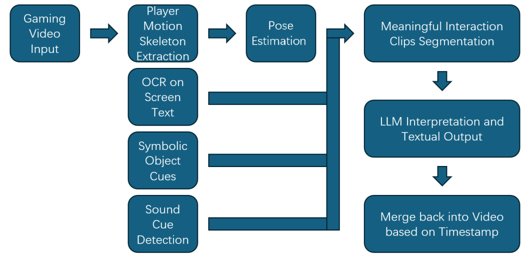

Figure 2: Overview of the proposed multimodal pipeline. Gameplay videos are processed to extract motion (via pose estimation), along with OCR-based text, object cues, and audio signals. These multimodal cues are used to segment interaction clips, which are then interpreted by an LLM to infer player intent. The results are mapped back to the video timeline as timestamp-aligned annotations.

For each segmented interaction clip, we construct a structured representation based on the previously extracted signals, including timestamped gesture labels, textual cues, and audio transcripts. We then supply the LLM with this temporally aligned, symbolic description of player behavior. The model then performs high-level reasoning over these inputs to infer player intent such as friendliness, surrender, cooperation, or resource offering, conditioned on predefined interpretation guidelines.

In the future, we plan to add multimodal cue extraction using Paddle v5 [Cui et al. 2025] and Whisper speech to text [Radford et al. 2023], interaction segmentation, and output reconstruction.

2.2 Preliminary Results:

Figure 1 shows example outputs from a sample gameplay video, where multiple gesture types (e.g., prone, crouch, pointing, and salute-like actions) are detected and aligned with corresponding timestamps. Testing on sample clips suggests that the system can reliably identify a subset of gestures, particularly those with distinctive joint configurations or repeated motion patterns. However, the current system is less reliable for gestures whose meaning depends on non-skeletal information.

Table 1 summarizes the current detectability of representative gestures in the prototype. We present them as a preliminary diagnostic of which parts of the gesture taxonomy are already supported by motion-based analysis and which require multimodal support.

3 Discussion

Emergent gestures in Delta Force are not only practical signals for avoiding combat; they also reveal how players negotiate trust, risk, friendliness, submission, temporary alliance, and possible betrayal under uncertainty. Our preliminary exploration highlights that a

Table 1: Taxonomy of non-hostile gestures, inferred meanings, and detectability unless not yet implemented (NYI).

Category	Gesture	Interpretation	Detectability
Atomic	Carrying a box (cannot use weapon)	Signals non-hostility by disabling combat capability.	NYI
	Holding "Happy Birthday" sign	Harmless emote with sound cue; indicates friendly intent.	Weak
	Kneeling pose	Explicit surrender or plea for mercy.	Strong
	Voice line "Happy New Year"	Announces presence to avoid being mistaken as a threat.	Weak
	Salute	Shows respect or gratitude.	Moderate
	Placing a container on the ground	Implies begging or requesting resources.	NYI
	Pointing at an item	Indicates offering or directing attention to an item.	Moderate to strong
Composite	Shaking left and right	Non-hostile signal; prevents precise aiming.	Moderate
	Repeated prone movement	Similar to bowing; expresses submission or friendliness.	Strong
	Repeated spinning	Harmless behavior; reduces threat perception.	Moderate
	Repeated shield + jump	Highly visible signal; defensive and non-aggressive.	NYI
	Dropping equipment	Fully disarms oneself to signal harmlessness.	Weak
	Rhythmic shooting pattern	Community-established signal indicating surrender or intent to cooperate.	NYI

substantial portion of player interaction in extraction-based multiplayer games is mediated through informal, community-driven non-verbal signals. These signals are not explicitly designed by the game system but emerge from shared player practices and risk–reward dynamics. This suggests that player communication in such environments extends beyond predefined mechanics and forms an implicit "language" that is learned socially.

One key motivation of this work is to make these implicit communication patterns more accessible. For experienced players, these signals are often intuitive, but for newcomers or external viewers, they can be difficult to interpret. A system that can automatically detect and explain such behaviors may support onboarding, reduce misinterpretation, and enhance spectatorship. More broadly, this points to the potential of AI-assisted interpretation systems in making tacit social knowledge visible in interactive environments.

A central challenge in interpreting non-verbal communication in extraction shooters is that signals are not always sincere. Because players may benefit from lowering an opponent's guard, gestures that conventionally indicate friendliness, surrender, or cooperation can also be used deceptively. Future versions of the pipeline could incorporate subsequent actions, such as whether a player attacks after signaling surrender or whether both parties disengage peacefully, to distinguish stable cooperation from deceptive signaling.

This work is currently at a preliminary stage and has several limitations. Future work will focus on expanding the gesture taxonomy, improving robustness of detection, and conducting user studies to evaluate how well the inferred interpretations align with player understanding.

4 Conclusion

This is an initial framework for studying non-verbal communication in Delta Force through multimodal AI analysis. Extraction shooter games provide a rich setting for examining how players communicate. Our preliminary taxonomy and prototype show that some gestures can be detected from gameplay video, while others require additional contextual information. This work offers a foundation for studying social behavior and emergent communication practices in multiplayer games.

Acknowledgments

We thank Peter Xenopolous and Boyuan Li for their advice. ChatGPT Deep Research was used for related work, Github Copilot Claude Sonnet 4.6 for programming and debugging, and ChatGPT for gaming content analysis and table formatting.

References

MMPose Contributors. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. PaddleOCR 3.0 Technical Report. arXiv:2507.05595 [cs.CV] https://arxiv.org/abs/2507.05595

Emanuela Guglielmi, Simone Scalabrino, Gabriele Bavota, and Rocco Oliveto. 2023. Using Gameplay Videos for Detecting Issues in Video Games. *Empirical Software Engineering* 28, 6, Article 136 (2023). doi:10.1007/s10664-023-10365-0

Yubo Kou and Xinning Gui. 2014. Playing with Strangers: Understanding Temporary Teams in League of Legends. In *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play*. 161–169. doi:10.1145/2658537.2658538

Alex Leavitt, Brian C. Keegan, and Joshua A. Clark. 2016. Ping to Win?: Non-Verbal Communication and Team Performance in Competitive Online Multiplayer Games. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4337–4350. doi:10.1145/2858036.2858132

Juhoon Lee, Seoyoung Kim, Yeon Su Park, Juho Kim, Jeong-woo Jang, and Joseph Seering. 2025. Less Talk, More Trust: Understanding Players' In-game Assessment of Communication Processes in League of Legends. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Article 412. doi:10.1145/3706598.3714226

Aaron May, Amber Choo, Carman Neustaedter, and Alissa Nicole Antle. 2013. A Mixed-Methods Evaluation of Nonverbal Communication in Collaborative Video Games. In *2013 IEEE International Games Innovation Conference*. 169–174. doi:10.1109/IGIC.2013.6659127

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.

Nemanja Rašajski, Chintan Trivedi, Konstantinos Makantasis, Antonios Liapis, and Georgios N. Yannakakis. 2024. BehAVE: Behaviour Alignment of Video Game Encodings. In *Computer Vision – ECCV 2024 Workshops*. 321–338. doi:10.1007/978-3-031-92387-6_23

Fredrik Rusk, Matilda Ståhl, and Nicholas Taylor. 2024. Callouts as a Coordinating Device in a Team-Based Networked First-Person Shooter Game. *Social Sciences & Humanities Open* 9 (2024), 100753. doi:10.1016/j.ssaho.2023.100753

Mohammad Reza Taesiri and Cor-Paul Bezemer. 2025. VideoGameBunny: Towards Vision Assistants for Video Games. In *Proceedings of the Winter Conference on Applications of Computer Vision*. 1403–1413. doi:10.1109/WACV61041.2025.00144

Evelyn T. S. Tan, Katja Rogers, Lennart E. Nacke, Anders Drachen, and Alex R. Wade. 2022. Communication Sequences Indicate Team Cohesion: A Mixed-Methods Study of Ad Hoc League of Legends Teams. *Proceedings of the ACM on Human-Computer Interaction* 6, CHI PLAY, Article 225 (2022). doi:10.1145/3549488

Anthony Tang, Jonathan Massey, Nelson Wong, Derek Reilly, and W. Keith Edwards. 2012. Verbal Coordination in First Person Shooter Games. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. 579–582. doi:10.1145/2145204.2145292

Jason Wuertz, Scott Bateman, and Anthony Tang. 2017. Why Players use Pings and Annotations in Dota 2. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. doi:10.1145/3025453.3025967

Keyang Zheng and Rosta Farzan. 2023. Understanding Player's Gesture-Based Communicative Behavior in MOBA Games. *Proceedings of the ACM on Human-Computer Interaction* 7, CHI PLAY (2023), 1068–1090. doi:10.1145/3611061

Keyang Zheng, Ben Stein, and Rosta Farzan. 2023. Use Ping Wisely: A Study of Team Communication and Performance under Lean Affordance. *ACM Transactions on Social Computing* 5, 1–4 (2023). doi:10.1145/3557022