

A General-Purpose Vision and Audio Feedback Pipeline for Camera-Observed Physical Puzzles in K–5 Learning

Aranis Das, Shubhendu Das

aranis.das@stonybrook.edu, shubhendu@steamworksstudio.com

Abstract— Hands-on manipulatives help K–5 learners build spatial reasoning and foundational STEM concepts, but most physical puzzles provide limited real-time guidance, assessment, or accessibility support. We present a general-purpose computer vision pipeline that recognizes incremental state changes in camera-observed physical puzzles and drives low-latency spoken feedback (audio talkback) without requiring fiducial markers on every piece or a high-compute cloud vision model. Our approach converts each camera frame into a canonical top-down coordinate system using lightweight corner-anchor detection, then performs robust change detection against an adaptive baseline using gradient-domain differencing under global drift checks. Puzzle states are represented as polygonal regions in canonical space, enabling fast per-region evidence scoring to infer which piece was placed. Temporal confirmation and hysteresis reduce false triggers from partial placements, occlusions, and lighting changes. The system supports a modular feedback loop for tutoring behaviors (e.g., prompts, hints, verification, and error recovery) and is designed to transfer across puzzle geometries by swapping only the state definition file and a small set of anchor templates. We describe the pipeline, design tradeoffs for embedded and classroom settings, and an evaluation plan targeting both recognition reliability and learning-facing metrics (engagement, autonomy, and error recovery time).

Index Terms—Audio feedback, camera-based interaction, computer vision, embedded learning systems, manipulatives, tangible user interfaces

I. INTRODUCTION

Physical puzzles and manipulatives are widely used in K–5 classrooms and at-home learning to develop spatial reasoning, categorization, and early geography, math skills and reading. However, physical kits typically lack real-time feedback: learners may not know whether a piece is correctly placed, what to try next, or how to recover from mistakes without adult intervention. This gap is especially costly for learners who benefit from frequent scaffolding or alternative modalities such as spoken guidance.

Beyond geography, the proposed system generalizes to any learning puzzle where the spatial placement of a piece encodes semantic meaning. Examples include letter puzzles (placement of characters in an alphabet board), number puzzles (numerical ordering or arithmetic layouts), and anatomy puzzles (correct placement of anatomical regions). In these settings, the system can emit contextual spoken feedback such as “That is the letter A — A for Apple!” or “This region is the frontal lobe,

which is responsible for executive function,” directly coupling placement with semantic explanation.

Traditional physical puzzles are rather passive as they confirm correctness only after completion and provide no real-time semantic or corrective feedback. This limits opportunities for scaffolding, error recovery, and active engagement during learning, particularly for early learners and self-guided environments. [1]–[3]

Computer vision (CV) enables physical puzzles to provide immediate guidance by observing learner actions and responding with spoken feedback. [4] Prior work in embodied and multimodal learning shows that coupling physical interaction with timely visual and auditory feedback can improve engagement, understanding, and recall, particularly for early learners. [3], [5] However, deploying such systems in low-cost classroom settings is challenging due to lighting variation, hand occlusion, motion blur, and camera drift.

We therefore present a general-purpose, low-latency vision pipeline for camera-observed learning puzzles that prioritizes embedded feasibility and transferability. By linking physical placement to immediate visual and spoken feedback, the system transforms static manipulatives into interactive tutoring artifacts that actively guide learners through placement, meaning, and recall.



Fig. 1: Example learning domains supported by the proposed system. Left: alphabet puzzle with letter placement and phonetic feedback. Center: number puzzle enabling numerical or arithmetic guidance. Right: anatomy puzzle illustrating semantic feedback tied to spatial brain regions.

Images adapted from commercial educational toys (Amazon product listing; Pearhead; Mushie)

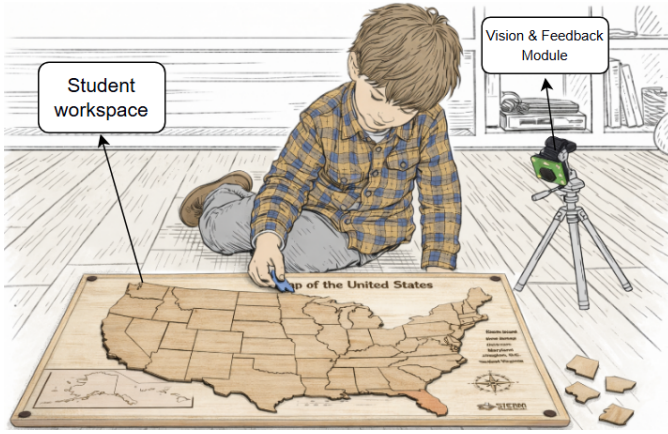


Fig. 2: Conceptual overview of the learning setup. A student interacts with a physical puzzle while a top-down vision system observes placement and provides immediate feedback.

II. RELATED WORK

A. Tangible learning and camera-observed manipulatives

A substantial body of work in learning sciences and human-computer interaction demonstrates that tangible manipulatives can improve engagement, spatial reasoning, and conceptual understanding, particularly for early learners and STEM education contexts [6], [7], [8]. Physical interaction has been shown to support embodied cognition, allowing learners to ground abstract concepts in sensorimotor experience [9], [10]. Prior studies further indicate that discovery-based and hands-on learning with manipulatives can outperform purely symbolic or screen-based instruction, especially when learners receive timely feedback [1], [2].

Camera-observed tangible systems extend these benefits by enabling automated assessment and guidance without instrumenting individual objects. Several systems have explored vision-based observation of learner actions to infer task progress or correctness [13], though many rely on controlled setups, fiducial markers, or high-end sensing infrastructure.

B. Embodied and multimodal learning

Embodied learning theory suggests that cognition is deeply linked to physical action and perception, and that learning environments which integrate movement, vision, and sound can improve retention and transfer [9], [14]. Multimodal feedback, particularly the combination of visual and auditory cues, has been shown to enhance learner engagement and conceptual clarity across a range of educational tasks [10], [15]. Tangible user interfaces that incorporate multiple sensory channels can further support meaning-making by reinforcing the relationship between action, location, and semantic content [6], [16].

C. Audio and multisensory feedback in educational interfaces

Audio feedback has been identified as a particularly effective modality for scaffolding and reinforcing learning, especially for young learners and self-guided environments [15],

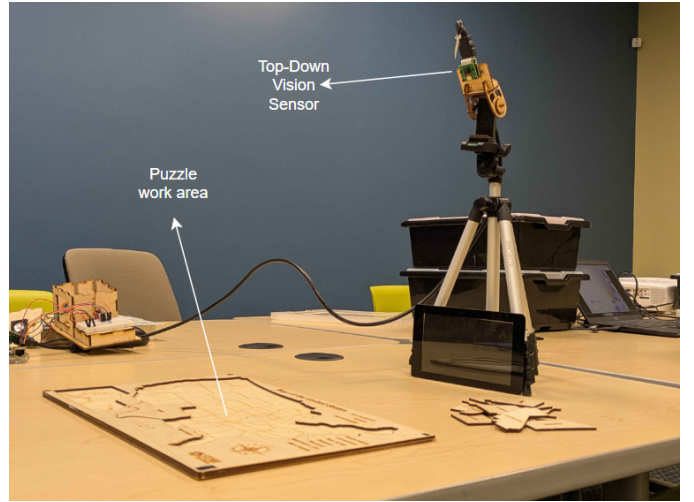


Fig. 3: Physical system setup used in this work. A top-down camera observes the puzzle workspace, enabling vision-based detection of piece placement under real-world conditions.

[17]. Prior work on multisensory and audio-tactile learning systems shows that spoken or auditory cues can improve task understanding, encourage exploration, and support recall when paired with physical interaction [16], [18]. These findings motivate systems that provide immediate spoken responses to learner actions, rather than relying solely on visual confirmation.

D. Marker-based versus markerless recognition

Many camera-based tangible systems employ fiducial markers or printed tags to achieve robust detection and tracking [13]. While effective, such approaches can reduce aesthetic quality, increase per-piece labeling overhead, and constrain puzzle design. Markerless approaches preserve the physical experience but often require strong appearance models, controlled backgrounds, or computationally expensive inference [19]. Recent work explores lightweight alternatives that combine global alignment with localized change detection to balance robustness and deployability in real-world learning settings [20].

E. Embedded vision and low-latency constraints

Embedded vision pipelines for learning technologies must operate under strict latency, power, and reliability constraints, particularly in classroom or at-home deployments [5]. Prior research highlights that consistent behavior, graceful handling of uncertainty, and recoverability from errors may be more important than maximizing recognition accuracy in educational contexts [1], [5]. These considerations motivate vision-based tutoring systems that emphasize drift detection, temporal confirmation, and low-latency feedback over heavy neural inference or cloud-dependent processing.

III. SYSTEM OVERVIEW

Fig. 4 shows the end-to-end architecture. A camera observes a puzzle board, then each frame is warped to a canonical

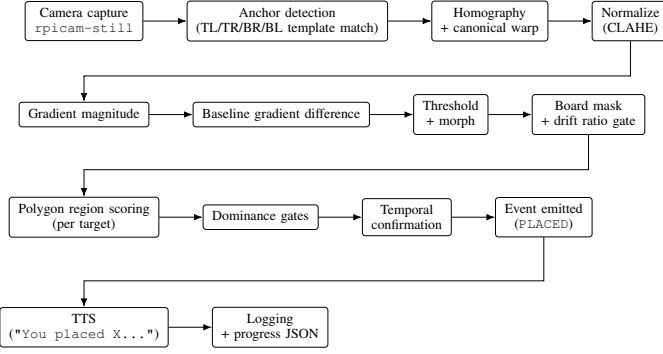


Fig. 4: System overview pipeline from capture to confirmation and feedback.

coordinate system using corner anchors. The CV module infers incremental placement events, then passes events to a tutoring policy that generates audio talkback (TTS) and logs interaction data for evaluation.

A. Design goals

We target three practical goals:

- 1) **Low latency:** feedback should feel immediate after a child places a piece.
- 2) **Robustness:** tolerate hand occlusions, lighting changes, and minor camera drift.
- 3) **Transferability:** adapt to new puzzles with minimal re-training and setup.

IV. METHOD: CANONICAL WARP + DRIFT-AWARE CHANGE INFERENCE

This section describes the core recognition method. The implementation is designed to run in a loop on a small computer (e.g., Raspberry Pi class devices) and can be applied to any puzzle that has (i) a roughly planar board and (ii) stable anchor locations.

Fig. 5 provides a qualitative pipeline from a single captured frame, showing the exact intermediate images produced by our implementation: (a) the raw camera view, (b) corner-anchor ROIs and detected template peaks, (c) the canonical top-down warp via homography, (d) local contrast normalization (CLAHE), (e) gradient-magnitude features used for change inference, and (f) polygon overlays in canonical space used for region scoring and feedback.

A. Anchor detection and canonical homography

Let I_t be the camera image at time t . We detect four corner anchors (TL, TR, BR, BL) using template matching within per-corner regions of interest, yielding image-space points p_k^{img} for $k \in \{TL, TR, BR, BL\}$. We compute a homography H_t that maps these points to fixed canonical coordinates p_k^{can} and warp the image:

$$I_t^{can} = \text{warp}(I_t, H_t). \quad (1)$$

This provides a stable coordinate system in which puzzle regions and masks are defined once and reused.

B. Gradient-domain differencing with adaptive baseline

Direct pixel differencing is sensitive to illumination changes. We instead compute a normalized grayscale image and use Sobel gradients to form a gradient magnitude image G_t . We maintain an exponential moving average baseline $\bar{G}_t$:

$$\bar{G}_t = (1 - \alpha)\bar{G}_{t-1} + \alpha G_t. \quad (2)$$

We compute absolute difference $D_t = |G_t - \bar{G}_t|$ and threshold it (Otsu or fixed threshold) to form an edge-like change mask M_t . Morphological operations remove speckle and stabilize edges.

C. Global drift rejection

Large global changes often correspond to hands covering the board, camera bumps, or major lighting shifts. We define a board mask B and compute a normalized change ratio:

$$r_t = \frac{\|M_t \wedge B\|_0}{\|B\|_0}. \quad (3)$$

If r_t exceeds a drift threshold, we skip placement inference for that frame and update the baseline aggressively to recover quickly once the scene stabilizes. We omit drift/occlusion examples from the progress strip (Fig. 5) for readability, but drift gating is active online and is critical for stability under hands and lighting changes. During development, we observed that transient occlusions and global illumination shifts could induce large gradient responses, motivating explicit drift detection and confirmation gating as a first-class design constraint.

D. Region scoring for placement inference

Each puzzle target (e.g., a state region) is represented by a polygon in canonical coordinates. We precompute a binary mask P_i for each target i . For each frame we compute a score:

$$s_i(t) = \|M_t \wedge P_i\|_0. \quad (4)$$

We compare scores to previous frame scores and take a *delta* $\Delta_i(t) = s_i(t) - s_i(t-1)$ to focus on incremental change. We select the best candidate by dominance:

$$i^* = \arg \max_i \Delta_i(t), \quad (5)$$

with absolute and relative gates (minimum delta, minimum pixels, dominance ratio, dominance margin).

E. Temporal confirmation and hysteresis

To reduce false positives, a candidate must satisfy gates for N consecutive frames before a placement event is emitted. Additionally, once a target is accepted as placed, it is excluded from further detection unless explicitly configured otherwise.

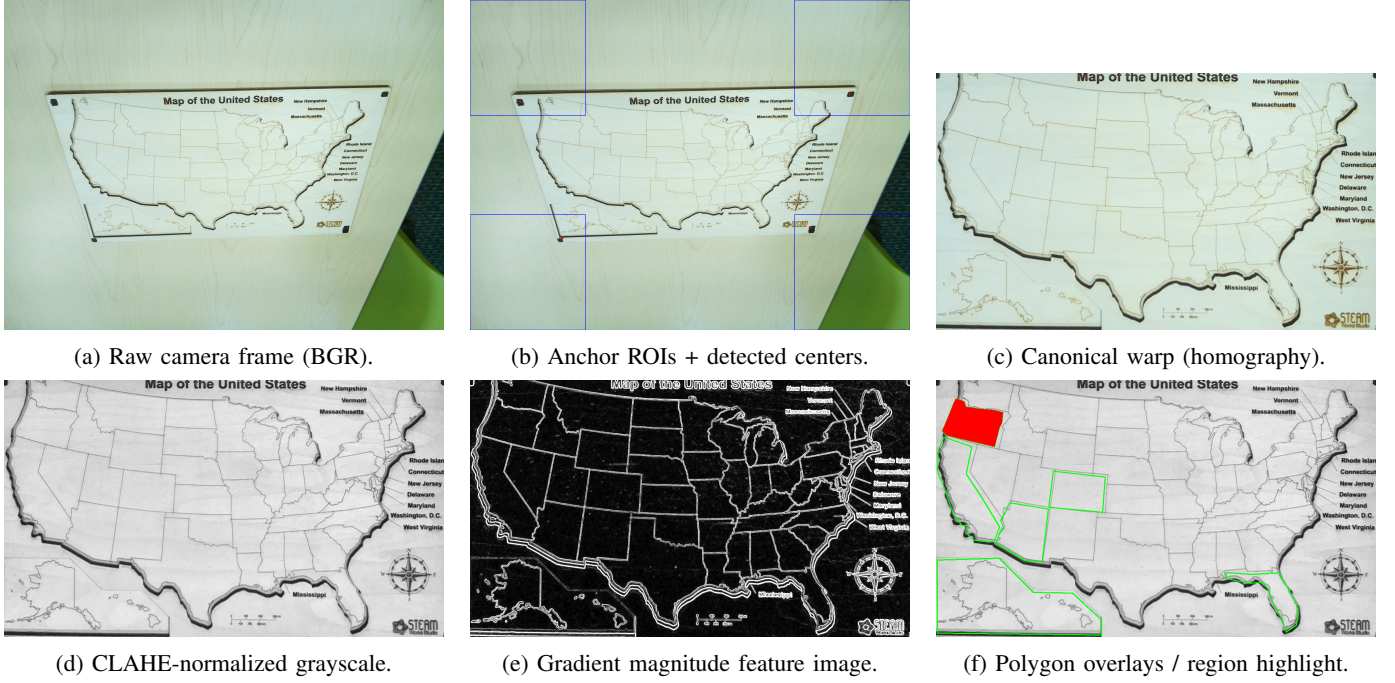


Fig. 5: Pipeline images generated from the same captured frame. The system detects corner anchors, warps the board into a canonical reference frame, normalizes local contrast, computes gradient-domain features for change inference, and overlays polygonal regions for downstream scoring and audio feedback. In our physical prototype, the corner anchors (TL, TR, BR, BL) were hand-drawn directly on the board using a permanent marker, demonstrating that the system does not rely on precise fiducials or printed markers for stable alignment.

TABLE I: Core parameters fixed across experiments and development runs.

Parameter	Value
Baseline EMA update (α)	0.18
Mid-frame baseline update (α_{mid})	0.06
Drift ratio threshold (r_{drift})	0.18
Maximum update ratio (r_{update})	0.08
Dominance ratio (ρ)	2.0
Dominance margin (pixels)	600
Temporal confirmation frames (N)	2
Minimum region pixels	250

V. EXPERIMENTS AND SYSTEM BEHAVIOR

The recognition module emits events such as `PLACED` (i), `UNCERTAIN`, and `DRIFT`. A tutoring policy consumes events and chooses audio responses:

- **Verification prompts:** ask the learner to confirm or re-seat a piece when confidence is low.
- **Content questions:** after correct placement, ask an educational question (capital, spelling, arithmetic, pattern reasoning).
- **Hints:** provide next-step suggestions based on remaining targets.
- **Recovery:** if drift persists, request that hands move away or the board be re-centered.

We evaluate recognition reliability, temporal stability, and system responsiveness under realistic classroom interactions. In a recorded end-to-end interaction, the time from physical

placement completion to spoken confirmation (“You placed *state*!”) was approximately 3.2 seconds, dominated by temporal confirmation and speech synthesis rather than per-frame visual processing.

Across logged development runs containing hand motion and lighting changes, only a small fraction of high-gradient frames resulted in confirmed placements due to dominance gating and temporal confirmation, indicating effective suppression of false triggers.

VI. LIMITATIONS AND DISCUSSION

The proposed system prioritizes robustness and explainability over per-frame responsiveness. As a result, confirmation latency is dominated by temporal consistency checks and audio synthesis rather than raw visual processing speed. While this design choice increases end-to-end delay, it substantially reduces false triggers under realistic classroom conditions involving hand motion, partial occlusion, and lighting variation.

The current implementation assumes a fixed board layout and relies on corner anchor visibility for canonical warping. Severe occlusions of multiple anchors or extreme camera displacement can temporarily prevent confirmation. In such cases, the system emits explicit `DRIFT` or `UNCERTAIN` events and requests user intervention (e.g., re-centering the board), rather than producing incorrect feedback. This behavior reflects an intentional design decision to fail safely and transparently.

Our evaluation is limited to development-time logs and a small number of end-to-end interaction recordings, rather than large-scale classroom deployments. While these logs demonstrate strong false-trigger suppression and stable placement confirmation, broader longitudinal studies are required to quantify learning outcomes, long-term engagement, and educator usability across diverse settings.

Finally, the tutoring policy currently operates on a fixed event schema. Although sufficient for geography-based tasks, more complex curricula (e.g., multi-step reasoning or adaptive difficulty) would benefit from tighter integration between recognition confidence, learner history, and pedagogical state.

VII. CONCLUSION

We presented a vision-based interactive tutoring system that links physical puzzle manipulation with real-time audio feedback using a robust, explainable recognition pipeline. By combining canonical warping, contrast normalization, gradient-based region scoring, dominance gating, and temporal confirmation, the system reliably detects correct placements while suppressing spurious triggers caused by motion, occlusion, and lighting variation.

Unlike approaches that emphasize per-frame classification accuracy, our design treats stability over time as the primary signal for correctness. This enables the system to operate effectively in unstructured, hands-on learning environments without requiring instrumented pieces, fiducial markers, or wearable sensors.

The resulting event-driven interface supports a range of tutoring behaviors, including verification prompts, content questions, hints, and recovery strategies, making the platform adaptable to multiple educational domains beyond geography. Future work will focus on classroom-scale deployment, richer learner modeling, and tighter integration between recognition confidence and adaptive instructional policies.

ACKNOWLEDGMENT

The authors thank Shubhendu Das, CEO of Steamworks Studio, for support, feedback, and access to prototyping resources. This research was conducted at Steamworks Studio in Princeton, New Jersey.

REFERENCES

- [1] K. J. Carbonneau, S. C. Marley, and J. P. Selig, "A meta-analysis of the efficacy of teaching mathematics with concrete manipulatives," *Journal of Educational Psychology*, vol. 105, no. 2, pp. 380–400, 2013.
- [2] E. V. Laski and R. S. Siegler, "Learning from number board games: You learn what you encode," *Developmental Psychology*, vol. 51, no. 7, pp. 853–864, 2015.
- [3] J. Hattie and H. Timperley, "The power of feedback," *Review of Educational Research*, vol. 77, no. 1, pp. 81–112, 2007.
- [4] H. Ishii and B. Ullmer, "Tangible bits: Towards seamless interfaces between people, bits and atoms," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1997, pp. 234–241.
- [5] D. S. Weisberg, K. Hirsh-Pasek, and R. M. Golinkoff, "Guided play: Where curricular goals meet a playful pedagogy," *Mind, Brain, and Education*, vol. 11, no. 3, pp. 142–152, 2017.
- [6] R. Moreno and R. E. Mayer, "Interactive multimodal learning environments," *Educational Psychology Review*, vol. 19, no. 3, pp. 309–326, 2007.
- [7] O. Zuckerman, S. Arida, and M. Resnick, "Extending tangible interfaces for education: Digital montessori-inspired manipulatives," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2005, pp. 859–868.
- [8] A. N. Antle, "Research opportunities: Embodied child–computer interaction," *International Journal of Child-Computer Interaction*, vol. 1, no. 1, pp. 30–36, 2013.
- [9] P. Dourish, "Where the action is: The foundations of embodied interaction," *MIT Press*, 2001.
- [10] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [11] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.
- [12] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*, P. S. Heckbert, Ed. Academic Press, 1994, pp. 474–485.
- [13] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.
- [14] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer, 2010.
- [15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," in *International Journal of Computer Vision*, vol. 60, no. 2, 2004, pp. 91–110.
- [16] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision and Image Understanding*, vol. 110, no. 3, 2008, pp. 346–359.
- [17] G. Bradski, "The opencv library," *Dr. Dobbs's Journal of Software Tools*, 2000.
- [18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," pp. 886–893, 2005.
- [19] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 511–518, 2001.
- [20] C. M. Bishop, "Pattern recognition and machine learning," *Springer*, 2006.