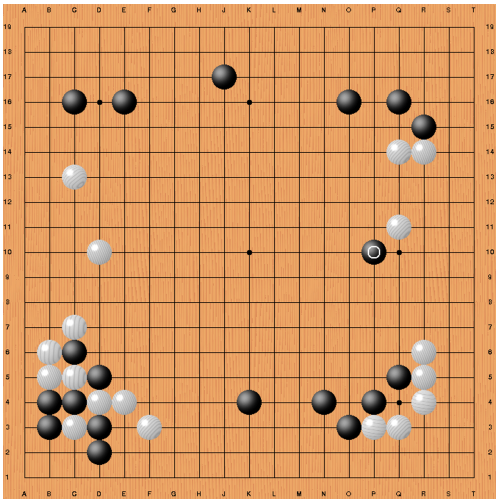
Visual Reasoning Models (VRMs)

Arun Kejariwal, Bhargav Bhushanam, Aayush Prakash, Siddharth Ramakrishnan, Saurabh Agarwal

Human thought and reasoning have long been the holy grail for advancing AI capabilities. AlphaGo's famous Move 37 - black stone on O10 (see the figure on the right) - became a landmark demonstration that AI models can "think." Subsequent iterations - AlphaGoZero, AlphaZero, MuZero - not only learned to master Go, Chess and Shogi from scratch without needing to be told the rules.



Progress has since extended from mastering perfect-information board games to, for example, complex scientific discovery [1, 2, 3, 4], programmatic optimization [5, 6], and formal mathematical reasoning [7, 8, 9]. Recent highlights include the Gemini Deep Think model winning the gold medal at IMO 2025 [10], Knuth solving an open graph decomposition conjecture via Claude [11], OpenAI disproving a conjecture associated with the unit distance problem in Discrete Geometry [12]. These milestones demonstrate significant reasoning depth. Yet, a critical limitation persists: the underlying core reasoning loops are strictly textual,[1] symbolic, and programmatic.[2] This gives rise to "textual drift" or "visual blindness". Humans, by contrast, reason in natively visual space - for instance, tracing tangled copper lines on a printed circuit board to find short circuits or sorting mixed inventory bins (e.g., finding the "red square box with a blue stripe" vs. "blue box with a red stripe") or when dealing with molecular structures, satellite maps, or geometric diagrams.[3]

Myth of Geometric Vision

Humans solve Euclidean geometry problems by inspecting drawn figures and using spatial intuition to construct auxiliary lines. In contrast, leading models such as AlphaGeometry and AlphaGeometry2, despite being proficient at solving geometry problems in IMO, do not process visual diagrams or raw images. Instead, they operate on a purely textual and symbolic intermediate representation:

1. *Autoformalization:* A Gemini-based language model translates natural language problem statements into symbolic predicates. The statement *"Let ABC be a triangle with a right angle at A"* is formalised into the text string: `triangle a b c; perp a b a c`.
2. *Symbolic Representation:* Primitives are represented as nodes in a directed graph, and relations (such as collinearity, congruence, perpendicularity, and parallelism) are mapped as typed edges. No visual canvas or pixel layout is presented to the networks.
3. *The Logical Search Loop:* A symbolic engine evaluates the graph using geometric axioms to infer new relations. If the proof stalls, the language model outputs a string of text specifying an auxiliary construction (e.g., `construct X as intersection of line a and circle w`).
4. *Numerical non-degeneracy checks:* A coordinate-tracking layer assigns float coordinates to points in Cartesian space, acting as an algebraic constraint solver to prune invalid configurations. This layer does not represent visual data.

---

[1] Chain-of-Thought (CoT), Tree-of-Thoughts (ToT) and Graph-of-Thoughts (GoT) structure intermediate reasoning as sequences, trees, and graphs, respectively. Kojima et al. [13] showed via Zero-shot-CoT that LLMs already possess latent multi-step reasoning abilities, elicited simply by prepending "Let's think step by step" to a prompt — without requiring task-specific examples. All of these approaches operate purely in the language domain and lack native support for visual or diagrammatic reasoning.

[2] Programmatic reasoning refers to systems that translate abstract queries into executable code, formal tactics, or mathematical constraint graphs, rather than relying on loose natural language.

[3] In fact, learning from infancy is visual-first. See Appendix A and Appendix B.

Visual representations, by contrast, intrinsically support perceptual inferences such as recognizing symmetry, enclosure, or containment. The visual-first reasoning space comprises four functional meta-categories:

- *Spatiotemporal Object Grounding & Tracking:* Understanding changing spatial layouts, dynamic occlusions, and object identities over time guided by implicit temporal constraints.
- *Visual Anomaly Reasoning:* Identifying deviations from a learned "nominal distribution" (normal state), predicting the cause of the anomaly, and verifying the spatial location.
- *Physical & Kinematic Scene Understanding:* Reasoning about the physical laws of the universe - measuring size, velocity, torque, mass, and predicting the trajectory of dynamic structures.
- *Egocentric Wearable Proactive Planning:* Mapping first-person continuous video streams to active daily routines, isolating relevant interaction states, and timing interventions.

Visual puzzles

Spot-the-difference puzzles are a classic challenge: the logic is embedded entirely in low-level pixel discrepancies rather than any abstract rulebook. Vision-Language Models (VLMs) cannot solve them in a single, feed-forward pass. ForeSight interleaves textual reasoning steps with low-level visual tools and mask-based feedback, allowing the model to re-examine fine-grained visual features and update its logic when a discrepancy is found. MVoT takes a different approach: rather than translating images into text, it generates visual thought traces interleaved with verbal ones, “sketching” reasoning directly over the layout and reducing spatial translation errors.

Despite this progress, current models still fail on surprisingly simple cases. For the prompt “How many rabbits?” applied to an image with five rabbits on the right (source: page 17 of Qwen Look Again), Gemini 3.5 Flash Extended, GPT-5.5, and Claude Opus 4.8 Max all incorrectly reported 4. In general, current models are susceptible to surface properties, camera pose, illumination imbalance, occlusion etc.

VRMs

When solving a complex multi-step reasoning problem containing visual evidence, a standard Vision Language Model (VLM) often suffers from *visual decay*: as it generates longer, step-by-step text responses, its attention to the original visual tokens diminishes rapidly, leading to logical hallucinations. VRMs address this issue through:

- *Visual Reflection:* A specialized agent constructs vision-centered reasoning data for supervised fine-tuning (SFT) cold-starts and applies a visual-attention-based reward model during RL to penalize text-only shortcuts.
- *Visual-Textual Interleaving:* Frameworks such as *VTI-CoT* actively interleave textual reasoning steps with focused visual segment reviews, keeping spatial and contextual evidence aligned with generated thoughts. (See Appendix C for why VRMs retain a core textual component.)
- *Low-Level Visual Tools:* Low-level visual operators and mask-based feedback are incorporated directly into the reasoning process, forcing visual re-examination when intermediate claims fail validation.

Large-scale generative video models (e.g., Veo 3, Sora 2) have demonstrated reasoning directly within visual and physical dimensions. Rather than translating images into text tokens, these models use a Chain-of-Frames (CoF) mechanism: to solve spatial or physical puzzles—navigating a maze, simulating gravity, predicting Jenga collisions—the model generates a video rollout, reasoning frame-by-frame

through pixel manipulation.[4] Textual output serves only to translate the visually simulated outcome into human-readable form.

Pixel-space vs. Abstract Latent Space

VRM designers choose between two paradigms for intermediate reasoning: Pixel-Space (materializing physical, visual outputs) or Abstract Latent Space (keeping reasoning within hidden embedding manifolds). The table below provides a comparative overview.

Dimension	**Pixel-Space Reasoning**	**Abstract Latent Space Reasoning**
Core Mechanic	Intermediate "thoughts" are generated as raw pixels (e.g., drawing sketches, rendering layout code, creating physical masks).	Intermediate thoughts are processed entirely as continuous high-dimensional vector embeddings within the transformer layers.
Pros	*High Interpretability:* Humans can visually trace the model's logic. *Stable Grounding:* Physically rendering a scene prevents structural drift over long reasoning turns.	*Extreme Efficiency:* Bypasses the heavy computational cost of decoding and rendering photorealistic pixels. *Non-Linear Thinking:* Can hold multiple parallel/branching hypotheses without forcing immediate serialization.
Cons	*The Computational Tax:* Significant GPU waste is spent synthesizing highly detailed, photorealistic pixels that are irrelevant to the actual logical task.	*Black-Box Limitation:* Representations are entirely opaque and uninterpretable. *Collapse Risk:* Highly prone to representational collapse where vectors lose diversity.

In a traditional VLM, a single reasoning step follows the following pipeline:

$$\text{Visual Input} \xrightarrow{\text{Encoder}} v_i \xrightarrow{\text{Self-Attention}} h_t \xrightarrow{\text{Projection } W} \text{Logits} \xrightarrow{\text{Softmax}} \text{Discrete Token } y_t$$

where, $h_t \in \mathbb{R}^d$ is the model's last hidden state (typically, 4096-dimensional), and $W \in \mathbb{R}^{d * |V|}$ is the language projection matrix that maps it to vocabulary space |V|. In a latent visual reasoning model, the projection matrix W and the softmax step are *bypassed entirely.*

- *The Feedback Loop:* Instead of mapping h_t to a word (e.g., "behind"), the raw hidden state is fed directly back into the transformer's input layer as the next embedding.
- *Multimodal Interleaving:* The next reasoning block receives a mixed sequence of original visual tokens $v_1, v_2, \ldots, v_n$ and the newly generated continuous "thought tokens" h_t, h_{t+1}, ...
- *Attention Mechanic:* Attention heads compute similarity between abstract thought vectors (h_t) and spatial visual features (v_i) across layers, steering the thought vector based on edges, lighting gradients, or movement vectors.

The primary cognitive advantage of latent-space reasoning is the prevention of reasoning path collapse:

- Natural language is sequential and deterministic: committing to a word commits to a single logical path. A dead end (e.g., misidentifying an occlusion) cascades into further errors.
- A continuous h_t can represent a superposition of multiple alternative paths simultaneously.
- When an object disappears behind a screen, h_t need not immediately decide what happened next—the vector occupies a broad region of latent space encoding all possibilities in parallel (a

breadth-first search in vector space). As subsequent frames arrive, attention updates progressively narrow the trajectory.

To prevent representation collapse - continuous latent states drifting into uninterpretable noise - models employ Latent Visual Reasoning (LVR):

- During training, a joint objective optimizes h_t with both a language head (predicting the next text token) and a parallel latent projection head trained to reconstruct visual tokens v_{ROI} for the region of interest:

$$\mathcal{L} = \mathcal{L}_{\text{next-token}} + \lambda \mathcal{L}_{\text{reconstruction}}(h_t, v_{\text{ROI}})$$

- The reconstruction constraint forces h_t to remain aligned with physical, spatial features (edges, velocity, contrast) extracted by vision encoders such as DINOv2 or CLIP.
- During multi-turn latent exchanges, one approach is to enforce a differentiable constraint requiring the original query to be mathematically recoverable from the latent state at any time. If alignment drifts, the reconstruction loss spikes and forces a trajectory correction.

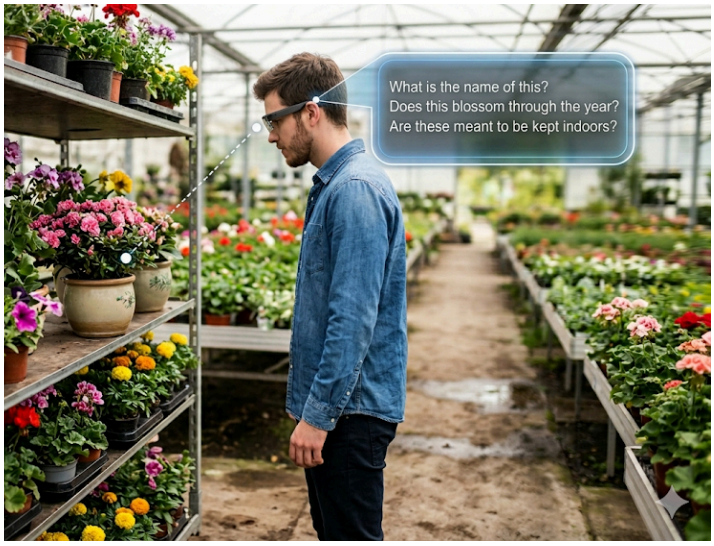
(Live) Video

Auto-formalizing a live video stream into a DSL in real time is not viable. Effective visual reasoning over video requires:

- *Persistent 3D Spatial Memory:* A persistent dual-memory module (working and episodic) maintains a view-consistent, 3D-aware representation purely from 2D video inputs, resolving camera pose shifts and maintaining object permanence during long occlusions.
- *Dynamic Occlusion Handling:* Models actively evaluate occlusion scores (leveraging tools such as SAM2) to filter keyframes, track boundaries, and propagate visual masks across highly dynamic sequences.
- *Spatiotemporal Reinforcement:* Tracking an actor through occlusion and low lighting requires a sequential decision process. Reinforcement-driven models use spatiotemporal rewards to train the model to "think before it segments," identifying exactly which frames contain the true visual evidence despite distortions.

Query-based vs. Query-free

In query-based egocentric settings, users interact directly via smart glasses, as illustrated on the right. Emerging multimodal agent frameworks bridge real-time perception, personal memory, and tool execution. Examples include:



- VisionClaw: *"Check reviews for this product. If there are more than 30 reviews, the combined 5- and 4-star score exceeds 85%, and the price is competitive on Amazon, add it to my cart."*
- SpatialClaw: *"What is the distance between the heater and the door, measured from the closest point of each?"*
- VisualClaw: *"The milk we usually buy for kids has run out—which type on the shelf is closest?"*

In query-free scenarios - e.g., a security camera detecting an intruder at 1 AM or smart glasses reminding you to grab an umbrella - no user prompt initiates the interaction. The role of the user's prompt is replaced by a persistent, background system

instruction (e.g., "Act as an ambient security guard. Continuously describe the video stream, check for anomalous human-object interactions, and evaluate safety risks"). Instead of running a one-off inference, the model runs a continuous loop. To prevent "textual drift" and ensure robust detection, the model evaluates live scene dynamics entirely in a visual latent space, monitoring a running spatiotemporal embedding of the environment. Instead of translating video clips into intermediate text strings which are prone to losing fine-grained contextual anomalies, the model measures deviations directly against a learned "nominal distribution". If the semantic distance between the current visual embedding state and a learned normal is too high, the system flags a deviation.

Specialist tools and visual critics serve as gatekeepers, pruning false positives before alerting the user. Rather than feeding raw, continuous video to a multi-modal LLM (MLLM), the exact frames where "something changed" are isolated (e.g., a person entering the camera frame) and fed to the model. If the model generates a potential alert - "Suspicious movement detected near the window" - a Visual Attention Critic audits the model's internal attention maps to verify that the model is grounded in the suspicious object, not reacting to textual bias or background noise. If the grounding is weak, the alert is suppressed. Alternatively, a mask-based critic temporarily masks the suspected region and asks the reasoning model whether the surrounding context still supports the anomaly claim, enforcing visual reflection before interrupting the user.

Models such as ROMA (Real-time Omni-Multimodal Assistant) process continuous environmental inputs as time-aligned multimodal units. It aligns dense audio inputs with discrete video frames to resolve granularity mismatches. Its internal language decoder continuously generates potential proactive actions. If its internal planning head determines that a high-value suggestion is strongly aligned with the immediate visual context (e.g., seeing grey clouds + pulling an external weather forecast tool), it proactively initiates a spoken-text alert.

The background system instruction does not require separate custom models for each use case. Instead, a single generalist VLM is configured into a specialized VRM through the following three mechanisms:

- *System Prompt Dynamic Re-Tasking:* The core VLM remains completely frozen. Swapping a system prompt from *"You are an industrial safety officer"* to *"You are a pediatric baby monitor"* changes the model's inference behavior without modifying any weights.
- *In-Context Exemplar Learning (ICL):* Rather than fine-tuning on labeled examples, the VLM receives 2–3 visual exemplars of "normal" and "anomaly" states directly in its context window, performing few-shot analogical reasoning in latent space.
- *Parameter-Efficient Adaptation (PEFT):* For sensitive industrial or clinical domains, researchers freeze the core VLM and train lightweight LoRA (Low-Rank Adaptation) adapters. For example, in AnomalyVFM, any frozen vision foundation model (such as DINOv2) becomes a zero-shot anomaly detector with a small feature adapter and a confidence-weighted pixel decoder.

Finally, in query-free scenarios, when operating in an abstract latent space, "reflection" becomes a constrained optimization process where the model's internal representation is continuously audited against fidelity anchors:

- *Self-Referential Consistency:* The previous hidden states act as the "prompt." If h_t contradicts h_{t-1} in terms of semantic vector stability, the model treats this as an error and adjusts the vector.
- *Manifold Smoothness:* The model optimizes for a smooth, differentiable transition across the latent manifold. Stochastic noise in the path degrades the final output.
- *Predictive Attractors:* The model predicts its own future latent state across a horizon; deviations from the task objective trigger a corrective latent update.
- *Contrastive Grounding:* The current latent h_t is compared against a cached representation of the original query, maintaining alignment with the query's intent.

Support for native visual-first reasoning is essential for building robust systems. However, in domains such as embodied AI, additional modalities must be incorporated for high-fidelity reasoning and follow-up action (see Appendix D).

Appendix

A. Pre-verbal intelligence

Infants as young as four months do not segment their visual field using language or Gestalt principles. Instead, they organize visual arrays into distinct physical bodies by analyzing surface arrangements and spatial-temporal paths of motion. This analysis is governed by four core physical constraints: cohesion (objects move as connected, bounded wholes), boundedness (distinct objects maintain separate boundaries and do not merge), rigidity (objects tend to maintain their shape and size over motion), and contact (objects only influence each other's trajectories upon direct physical touch). These principles demonstrate that a scalable, visual-first physical deduction system is hardwired into human cognition prior to linguistic development.

B. Early works in psychology on #VisualFirst learning

The following table throws light on the early works from different branches of psychology that proposed visual-first representation, especially in the context of how babies/infants learn.

	Key Contribution
The Child's Conception of Space (Piaget & Inhelder, 1956)	Spatial reasoning develops in a topological sequence (proximity, separation, order, enclosure) before transitioning to Euclidean frameworks - a native visual-spatial logic that precedes linguistic thought.
Realization of a geometry-theorem proving machine (Gelernter, 1959)	Demonstrated that a visual-like spatial model can effectively constrain symbolic reasoning, establishing that geometric "intuition" is computationally superior to blind logical deduction.
The Course of Cognitive Growth (Bruner, 1964)	Proposed three modes of representation: enactive (action-based), iconic (image-based), and symbolic (language-based). The iconic mode supports non-verbal reasoning before linguistic structures are acquired.
Visual Thinking (Arnheim, 1969)	Positioned visual perception as the primary arena of intelligence, showing that abstract concepts are fundamentally structured as spatial configurations.

Social Learning Theory (Bandura, 1977)	Demonstrated that pure observational learning—watching a model's behavior and its consequences—is sufficient for acquiring and generalizing complex novel behaviors.
Problem-solving with diagrammatic representations (Funt, 1980)	Argued that diagrammatic representations simplify physical reasoning by shifting the computational burden to parallel visual observations, avoiding complex text-based deductions.
Why a Diagram is (Sometimes) Worth Ten Thousand Words (Larkin & Simon, 1987)	Proved that visual information indexing inherently reduces search complexity, making it a highly scalable cognitive format.
Principles of Object Perception (Spelke, 1990)	Demonstrated that spatial coordinate tracking and physical deduction are deeply integrated into human visual processing—a primary evolutionary substrate.

C. Textual scaffolding of VRMs

VRMs retain a textual component for several structural reasons:

- Continuous spaces excel at representing texture, shape, and spatial gradients, but they struggle with strict compositional logic. Language provides a discrete, rule-based symbolic scaffolding. Concepts such as negation ("do not select the red block"), conditionals ("if X is present, then Y is true"), and quantifiers ("are all shapes circles?") are naturally encoded in natural language or code, allowing the model to perform rigorous, step-by-step logical deduction over perceived visual elements.
- Language acts as an executive planner for multi-step visual reasoning. In VALOR, for example, the language model decomposes a high-level spatial query (e.g., "Find the tool that was used after the box was opened") into discrete, sequential sub-steps: (i) locating the box, (ii) tracking the box's state over the video timeline, (iii) identifying the tool based on that temporal boundary. This chain of execution is naturally structured using natural language or Python-based tool-use scripts.
- Language is the only medium that can translate raw perceptual inputs (pixels) into verifiable formal representations (algebraic statements, 3D coordinate matrices, symbolic graphs) checkable by external logic verifiers.

D. Insufficiency of perception

Visual-first reasoning does not imply that perception alone is sufficient. This was demonstrated in the landmark "kitten carousel" experiment (Held and Hein, 1963): two groups of kittens received identical visual environments, but one walked freely while the other was carried passively. The passive kittens failed to develop normal depth perception, spatial orientation, or visually guided paw placement, demonstrating that perceptual systems require self-initiated, motor-driven feedback.

In child development, object manipulation catalyzes cognitive growth: multi-modal exploration (e.g., rotating an object while tracking it visually) teaches the brain to discover visual invariants such as object permanence, volume conservation, and spatial relationships. Multiple overlapping sensory systems—vision, audition, touch, proprioception—interact in degenerate ways to educate the cognitive system.

The same principle applies to embodied AI: video can teach an agent what to do conceptually, but cannot convey physical feedback—touch, pressure, and the precise force needed to peel a fish fillet or grasp a fragile egg.