Outstanding Questions

Luiz :

Some outstanding questions about the topics (i) "Biological Inspired Computer Vision" and (ii) "Active Foveated Reconstruction and Understanding of 3D Scenes".

1 - The Eye is not a Camera. Why the majority of research work in Computer Vision simply assumes a static pinhole camera model?

2 - Binocular stereo is a hard problem. Traditional methods are based on multiview geometry, while more recent methods exploit machine learned models from data. Both have been extensively studied by the computer science and engineering community. And yet, these dominant lines of research mostly disregard human stereoscopy. Why?

3 - Does it make sense to teach a PhD level course combining the two topics above? Why yes or not? What would be the goals?

4 - These two topics constitute promising scientific research directions that are worthwhile pursuing? What are the challenges and potential benefits?

5 - Can the research results from these two topics lead to practical applications? If so, in which areas? What innovations can be envisioned? Those would be based on which fundamental principles?

Gemini 3.6 Flash

1. Why Computer Vision Relies on Static Pinhole Camera Models

The dominance of the static pinhole camera model in computer vision is driven by mathematical simplicity, hardware standardization, and optimization efficiency:

- **Mathematical Tractability:** The pinhole model maps 3D points (X, Y, Z) to 2D coordinates (x, y) via linear perspective projection in projective space ($\mathbb{P}^2 and \mathbb{P}^3$). This linear matrix formulation enables explicit algebraic solutions for core problems such as projective reconstruction, bundle adjustment, and epipolar geometry.
- **Hardware and Manufacturing Alignment:** CMOS/CCD image sensors, optical camera lenses, and standard displays are physically built around flat, uniform rectangular grids. Frame-based imaging natively produces dense, structured arrays of pixels that match standard computing architecture.
- **Computational Frameworks:** Early computer vision was built on classical vector calculus, signal processing, and matrix algebra—all of which operate efficiently on uniform 2D spatial grids. Modern deep learning architectures (e.g., Convolutional Neural Networks and Transformers) similarly rely on dense tensor operations optimized for parallel processing GPUs.
- **Biological Complexity Trade-off:** Biological eyes feature non-uniform photoreceptor distributions (fovea vs. periphery), saccadic movement, non-linear optical distortion, continuous-time asynchronous signal output (retinal ganglion spikes), and active biomechanical control. Modeling these mechanisms introduces non-stationary, non-linear equations and asynchronous data representations that complicate standard algorithmic design.

2. Disregard of Human Stereoscopy in Dominant Stereo Research

Both traditional multiview geometry and modern data-driven stereo methods favor frame-based geometry over the mechanics of biological binocular vision due to differing operational paradigms:

- **Static vs. Active Paradigms:** Classical and deep learning stereo approaches primarily solve *passive offline matching*—finding corresponding points across pre-captured, spatially rectified frame pairs. Human binocular vision is inherently an *active closed-loop system* driven by dynamic vergence, accommodation, continuous saccades, and foveal fixations.
- **Uniform Resolution vs. Foveation:** Traditional stereo matching relies on dense cost volumes computed over uniform-resolution pixel grids. Human stereoscopy operates with high spatial acuity only at the fovea (covering ~2° of the visual field), using peripheral vision primarily for coarse motion detection, context, and directing the next vergence movement.

- **Geometric Rectification Convenience:** Computer vision simplifies matching from a 2D search to a 1D horizontal line search by pre-rectifying stereo images onto a single image plane. Biological vision lacks a fixed, pre-rectified pixel plane; human eyes continuously adjust vergence angles and compute horizontal/vertical retinal disparities dynamically across non-uniform surfaces (Panum's fusional area).
- **Supervised Learning and Dataset Bias:** Deep learning stereo models (e.g., PSMNet, RAFT-Stereo) rely heavily on dense 2D ground-truth disparity maps generated via LIDAR or synthetic 3D scenes. Training on static, densely labeled pixel frames reinforces models optimized for global, uniform disparity estimation rather than selective visual attention and saccadic exploration.

3. Feasibility, Goals, and Structure of a PhD-Level Course

Teaching a PhD-level course combining **Biologically Inspired Computer Vision** and **Active Foveated 3D Scene Reconstruction** is both highly viable and academically timely. It addresses fundamental limitations in current AI, such as bandwidth bottlenecks, static frame constraints, and high computational costs.

Course Goals

1. **Cross-Disciplinary Grounding:** Equipping students with foundational concepts in neurobiology (retinal processing, oculomotor control, visual cortex pathways) alongside advanced 3D vision geometry (epipolar geometry, neural radiance fields, active sensing).
2. **Paradigm Shift to Active Perception:** Shifting focus from passive pixel-processing models to active perception loops where sensing, motor control, and scene understanding interact continuously.
3. **Next-Gen Hardware Familiarity:** Hands-on experience with asynchronous event-based cameras, gaze-tracking systems, pan-tilt-zoom visual mounts, and non-uniform log-polar representations.

4. Challenges and Potential Benefits of these Research Directions

Key Challenges

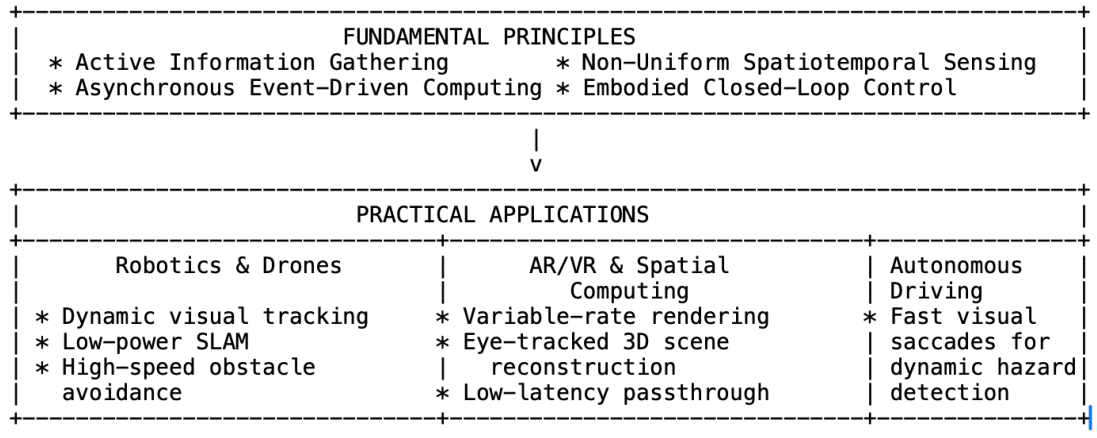
- **Asynchronous & Non-Uniform Computing:** Absence of standardized software pipelines for processing non-uniform (log-polar/foveated) visual representations and asynchronous event data compared to standard PyTorch/TensorFlow vision libraries.
- **Co-Optimization of Sensing and Perception:** Coupling mechanical or gaze control with perception algorithms requires solving complex, real-time reinforcement learning or optimal control problems under latency constraints.
- **Hardware and Dataset Scarcity:** Limited availability of standardized hardware setups combining active vergence mounts, high-speed foveated sensors, and precise 3D ground-truth capture systems for benchmarking.

Key Potential Benefits

- **Orders-of-Magnitude Compute and Bandwidth Reduction:** Foveated processing drastically lowers the pixel count required for high-precision scene understanding by processing high resolution only where necessary.
- **Ultra-Low Latency Reaction:** Event-based and active foveated sensing enables dynamic visual systems to react at millisecond scales to environmental changes without processing redundant background frames.
- **Active Embodied Intelligence:** Bridges the gap between static vision models and robotics, allowing autonomous agents to dynamically query their visual environment for task-relevant spatial information.

5. Practical Applications, Innovations, and Fundamental Principles

These research directions can yield practical breakthroughs across several domain applications:



Emerging Technical Innovations

1. **Neuromorphic Foveated SLAM:** Simultaneous Localization and Mapping algorithms that combine low-power neuromorphic event sensors with foveated active zoom mechanisms to track high-speed dynamic objects in complex 3D environments.
2. **Active Gaze-Directed Radiance Fields / 3D Gaussian Splatting:** Dynamic 3D scene representations reconstructed in real time by actively directing visual attention to regions of high geometric uncertainty, rather than uniformly capturing dense multi-view images.
3. **Ultra-Low-Power Edge Vision Chips:** Vision-on-chip architectures integrating event-based sensor arrays with on-sensor foveation networks, drastically cutting sensor-to-processor bandwidth requirements for edge computing.

Claude Chat, Opus 5.5 High

These five questions hang together. My short version: computer vision has treated the camera as a *recording instrument*. The eye is a *sampling instrument under a budget, steered by a controller*. Most of what follows comes from that difference.

1. The eye is not a camera. Why does CV assume a static pinhole?

The pinhole itself is not the problem. Optically, the eye is a decent central projection. The real differences are three: **sampling** (non-uniform, eccentricity-dependent, with roughly 100M+ photoreceptors feeding about 1M optic nerve fibres), **motion** (saccades, vergence, microsaccades, and Listing's law as a constraint on degrees of freedom), and **coupling to action** (where you look is chosen for what you need to know). The field kept the projection and dropped the other three. Several reasons explain that:

- **The instrument made the model.** CMOS sensors are uniform grids. Lenses and calibration are engineered so that a pinhole on a uniform lattice is literally true of the device. For engineers this was a description of the hardware, not an assumption.
- **Mathematical payoff.** Projective geometry on a uniform lattice gives epipolar geometry, the fundamental matrix, bundle adjustment, and 1D search after rectification. It also gives shift-invariance, so FFTs, pyramids and CNNs work. Non-uniform sampling breaks translation equivariance, which most of the toolbox depends on.
- **The problem was defined as interpreting given images.** From Roberts' blocks world through Marr to internet-photo collections, the input is a picture someone else took. Marr's separation of computational theory from implementation made it legitimate to ignore how the image was acquired. SLAM moves the camera, but moving is not the same as *active*: the trajectory isn't chosen for perception.
- **Benchmark culture.** Static datasets (Middlebury, KITTI, ImageNet) can only evaluate passive systems. Active vision needs closed loops, meaning simulators or physical rigs, and comparisons at equal budget. The 1987–1995 wave (Aloimonos, Bajcsy, Ballard, Brown) faded partly because of this, partly because pan-tilt heads were slow and expensive, and partly because the geometric program (Faugeras, Hartley–Zisserman) and later deep learning delivered more measurable progress.
- **The eye solves a constraint engineers mostly didn't have.** Foveation is a response to bandwidth and energy limits. When uniform high-resolution sensing and compute are cheap, foveation has nothing to buy. Your own finding states this precisely: foveal weighting on a uniformly sampled sensor is strictly lossy. Once the uniform samples exist, foveation can only discard information. Its benefit lives entirely in the budget: what you don't sample, transmit or compute.

So the pinhole model is not a mistake. It is the optimum for a regime where sensing is cheap. The interesting question is where that regime ends: edge devices, XR headsets, robots on batteries, gigapixel scenes, latency-bound loops.

2. Why do the dominant stereo lines disregard human stereoscopy?

Partly because they solve a different problem with a different objective.

- **Different job.** Engineering stereo is used as a dense range-finder: a full depth map, every pixel, one shot, fixed parallel baseline. Human stereo gives hyperacute *relative* depth (around 10–20 arcsec) near fixation, coarse depth elsewhere, integrated over fixations and fused with monocular cues. Beyond a few metres it contributes little, and stereoblind people function well. An engineer looking at human stereo sees a near-space, fixation-local cue, not a range sensor.
- **Rectification erases the human geometry.** Verged, toed-in geometry (Vieth–Müller, horopter, Panum's area) is treated as a nuisance to be warped away. In humans, vertical disparity is a cue for viewing distance and eccentricity (Mayhew & Longuet-Higgins; Gårding et al.). After rectification it becomes an assumption rather than a measurement, so the information is gone by construction.
- **The metric rewards uniform coverage.** Bad-pixel percentages over the whole image cannot be won by a system that is precise only where it looks. A fixation-centric system loses the benchmark by design, even if it is better at the task.
- **It was tried, then left behind.** Early computational stereo was explicitly bio-inspired: Julesz's random-dot stereograms, Marr–Poggio cooperative stereo, coarse-to-fine schemes. It was abandoned when optimization methods (graph cuts, SGM) and then learned cost volumes won on benchmarks. Neuroscience offers phenomena and mechanisms, such as the disparity energy model, but rarely algorithms that compete.
- **Half-occlusions are treated as outliers.** In da Vinci stereopsis (Nakayama & Shimojo), unmatched regions carry depth and boundary information. Matchers treat them as holes to fill. This is exactly where your D-phase floor sits: about 8% of the area at depth edges and half-occlusions. It is also why classifying the foveal window (local half-occlusion versus a depth edge in the window) is a biologically motivated move rather than a heuristic.
- **Disciplinary separation.** Vision science asks how humans do it, CV asks how to make it work, and their venues and incentives rarely meet.

There is an irony here. Learned foundation stereo models now absorb monocular priors and effectively do cue combination, which converges with human vision by accident rather than design.

3. Does a PhD course combining both topics make sense?

Yes, but as **one** course with one thesis, not two topics side by side. Topic (ii) is the computational realization of topic (i). The unifying thesis would be that *vision is budgeted, active sampling*, with biology as a source of constraints and hypotheses rather than a template to copy.

Why it makes sense:

- Questioning the default model is a core research skill, and this material is absent from standard CV curricula.
- It is naturally interdisciplinary: geometry, sampling theory, estimation, control, psychophysics and rendering.
- It produces open problems at every level, from a matcher on non-uniform lattices to gaze policies, fixation-relative representations and edge-aware matching.
- With a closed-loop renderer such as your Blender engine, students can run real experiments without hardware.

The main risk is that it becomes a catalogue of biological facts or a biomimicry ideology. The antidote is a standing question for every biological feature: *under which constraint is this optimal, and does that constraint hold for our system?* Negative results are good teaching material here, and you have a documented sequence of them.

Possible goals:

1. **Conceptual:** vision as sampling and control under a budget. Distinguish "moving camera" from "active perception."
2. **Technical:** verged binocular geometry (horopter, Listing/Donders, vertical disparity); non-uniform sampling (cortical magnification, log-polar, spherical lattices); classical and learned stereo; information-based and coverage-based gaze policies; fusion over repeated looks.
3. **Methodological:** fair comparison at equal budget (rays, energy, latency), pre-registered decision rules, and reading psychophysics critically for computational content. Given your course plans, AI-assisted R&D practice fits here too.
4. **Practical:** each student or group leaves with a working closed-loop experiment and a short paper-style report, positive or negative.

4. Are these promising research directions? Challenges and benefits

They are promising, but I'd sharpen the claim. The payoff of active foveated vision shows up where **resolution or bandwidth is binding and the task is local**: manipulation, search, tracking, inspection, interaction. It does not show up clearly in whole-scene reconstruction, where uniform coverage is the objective. Your Phase C result points this way: at equal rays, spreading fixations bought accuracy and the information objective did not. That doesn't refute the direction. It suggests reconstruction is the hardest testbed in which to show the benefit.

Potential benefits

- Efficiency in sensing, transmission, compute and energy, which matters again for XR, edge robotics and very high-resolution scenes.

- Robustness through redundancy. In your D5, two agreeing looks served as a test where a single look could not, a genuinely active-vision gain.
- Fixation as an organizing principle: segmentation by zero disparity, small disparity search ranges, task-relative precision.
- Alignment with human perception wherever a human is in the loop (XR, displays, assistive tech).

Challenges

- **Evaluation:** no static benchmarks exist, so closed-loop simulation is required, along with a defined budget currency and baselines compared at equal budget.
- **A strong baseline:** uniform high-resolution capture plus foundation models is very good. You must find regimes where the budget actually binds.
- **Representation:** non-uniform lattices break CNN machinery. The options are reprojection to tangent planes, spherical or equivariant networks, or graph and point methods.
- **Policy:** gaze selection is a POMDP. Simple heuristics (coverage, inhibition of return) are hard to beat without a task-specific reward.
- **Hardware:** foveated sensors barely exist, and mechanical gaze is slow. Emulation by rendering, crops or multi-camera setups is necessary but not fully convincing.
- **Matching at edges and occlusions:** this is the classical floor, and exactly where biology suggests treating occlusion as signal.
- **Biology's own gaps:** much is known phenomenologically and little computationally. There is a constant risk of cargo-cult transfer.

5. Practical applications, innovations and principles

Areas and envisioned innovations

- **XR:** foveated rendering is already shipping in gaze-tracked headsets. The next steps are gaze-contingent reconstruction, where the wearer's own gaze is the policy and the headset's 3D capture spends effort where the user looks. Further steps include vergence-aware passthrough, varifocal displays that address the vergence–accommodation conflict, and stereo-comfort budgets derived from horopter and Panum's area.
- **Robotics and humanoids:** verging stereo heads that fixate the grasp target, get absolute distance from vergence, and get fine relative depth near fixation. Fixation-based figure–ground segmentation (Mishra & Aloimonos).
- **Autonomous vehicles and drones:** foveated high-resolution crops on distant or critical objects under compute and bandwidth limits, and event cameras as retina-like front ends.
- **3D capture and neural rendering:** foveated ray budgets and next-best-view acquisition for NeRF and Gaussian splatting. Your engine is essentially this.

- **Sensors and neuromorphic hardware:** log-polar or multi-resolution CMOS, event sensors, and on-sensor gaze control.
- **Perceptual media:** gaze-contingent compression and streaming for cloud rendering and video (in the line of Geisler & Perry).
- **Clinical and assistive:** binocular-vision assessment and therapy for amblyopia and strabismus via stereo displays, low-vision aids, and retinal prosthesis encoding.

Fundamental principles underneath

1. **Non-uniform sampling matched to a budget:** eccentricity-dependent resolution and cortical magnification.
2. **Perception as action:** sampling chosen in a closed loop to reduce uncertainty for a task.
3. **Fixation-relative coordinates:** disparity relative to the horopter, small search ranges, fixation as segmentation.
4. **Oculomotor kinematics as constraint:** Listing and Donders laws reduce degrees of freedom, and vergence and version carry metric information.
5. **Temporal integration and redundancy:** repeated looks as a test (consensus), not only as averaging.
6. **Multi-cue fusion with priors:** stereo is one cue among motion parallax, defocus, shading and learned priors.
7. **Occlusion as information:** half-occlusions mark boundaries and depth order rather than noise.

If useful, I can turn the course part into a draft syllabus.

ChatGPT, GPT 6 Extra High

From Biological Vision to Active 3D Scene Understanding

A scientific perspective on five fundamental questions, with implications for doctoral education, research, and technological innovation.

My central thesis

I believe these two topics are not only scientifically worthwhile, but potentially form the foundation of a coherent research program in computer vision.

However, I would frame the opportunity somewhat differently from simply making artificial vision more biologically realistic.

The central opportunity is to move from reconstructing the world from images to understanding the world through purposeful visual interaction.

This distinction changes the nature of the problem:

- Conventional computer vision asks: What can be inferred from the images we have?
- Active biological vision suggests a different question: What should we look at, how should we look at it, and what should we infer from the resulting observations?

The second formulation introduces perception, action, attention, uncertainty, memory, and task-specific representation into a unified computational framework.

It also connects naturally with your ongoing exploration of foveation, binocular vergence, saccadic surface exploration, and scene understanding.

There are important precedents. The active perception work of Ruzena Bajcsy (1988) and the Animate Vision framework of Dana Ballard (1991) explicitly challenged the passive-image paradigm. More recent research, including a 2025 review of bio-inspired computational imaging, demonstrates renewed interest in designing sensors and computational methods jointly.

The interesting question is why this perspective has not become more central to computer vision—and whether the technological and scientific circumstances are now sufficiently different to justify revisiting it.

1. The eye is not a camera. Why does computer vision predominantly assume a static pinhole camera?

1.1 The distinction is more fundamental than optical geometry

Strictly speaking, the eye does behave as an optical imaging device, and perspective projection is a useful approximation to retinal image formation.

But the human visual system is not a passive imaging system.

Consider the differences:

Conventional imaging model
Approximately uniform pixel sampling
Predetermined field of view
Frames acquired at fixed intervals
Fixed or externally controlled focus
Fixed binocular orientation or calibrated stereo rig
Images processed as measurements
Reconstruction often treated as the objective

The fundamental difference is not simply the shape of the retinal surface or the density of photoreceptors.

It is that the biological system controls its observations as part of the process of understanding the environment.

An image is therefore not the primary computational object. It is one observation within an ongoing perception-action process.

1.2 Why the pinhole model became dominant

I see four main reasons.

First: mathematical simplicity.

The pinhole model provides an elegant relationship between 3D points and 2D images:

$$\boldsymbol{x} \sim K[R \mid t]\boldsymbol{X}.$$

This allows us to formulate camera calibration, epipolar geometry, triangulation, structure from motion, and bundle adjustment within a rigorous mathematical framework. For much of the history of computer vision, this geometric foundation has been extraordinarily productive.

Second: the available hardware.

Digital cameras produced uniformly sampled rectangular images. It was natural to design algorithms around the measurements those devices delivered. Biological foveation would have introduced additional complications in optics, sensing, sampling, gaze control, synchronization, and image representation.

Third: computational convenience.

Uniform image arrays fit particularly well with convolution, parallel processing, GPUs, and later deep learning. The representation was determined partly by what computers could process efficiently.

Fourth: the research methodology.

Fixed images and standardized datasets support reproducible experiments and benchmarking. A system that decides where to look next creates a substantially harder experimental problem: performance depends on both the perception algorithm and the acquisition policy. In other words, the image-based formulation isolates a manageable portion of a much larger problem.

1.3 Was this a conceptual mistake?

I would say no, but it became a limiting abstraction when applied to general visual intelligence.

The pinhole camera remains entirely appropriate for geometric reconstruction, even inside an active biological-inspired system.

Indeed, I would retain projective geometry as a fundamental part of the proposed framework.

The problem arises when we implicitly equate:

$$\text{Visual perception} \equiv \text{Processing a set of images}.$$

A more complete formulation is:

$$\text{Visual perception} = \text{Sensing} + \text{Inference} + \text{Action} + \text{Memory}$$

These components are continuously coupled.

There is an important historical perspective here. Ballard argued in Animate Vision that gaze control and behavioral constraints can substantially reduce the computation required for visual tasks. The argument was not that projective geometry should be abandoned, but that the organization of the computational problem should change.

1.4 What I would take from this

I would not pursue a literal simulation of the human eye as the principal objective.

Instead, I would investigate three biological principles:

1. Nonuniform sensing: allocate spatial and temporal resolution selectively.
2. Active acquisition: choose observations in response to what is currently known or unknown.
3. Task-directed perception: acquire only the information needed to support the next perceptual or behavioral decision.

These principles are general enough to apply to artificial systems that do not physically resemble human eyes.

And that is an advantage: biology becomes a source of computational ideas rather than an engineering specification.

2. Binocular stereo: why have the dominant approaches largely disregarded human stereopsis?

This question is particularly interesting because the problem being solved by both natural and artificial systems is fundamentally related. Yet their computational objectives are different.

2.1 Conventional stereo and human stereopsis are not solving precisely the same task

In classical computer vision, binocular stereo typically seeks a disparity field:

$$d(x, y) = x_L - x_R$$

for appropriately rectified images, followed by depth recovery:

$$Z = \frac{fb}{d},$$

where f is focal length, b is the baseline, and d is disparity.

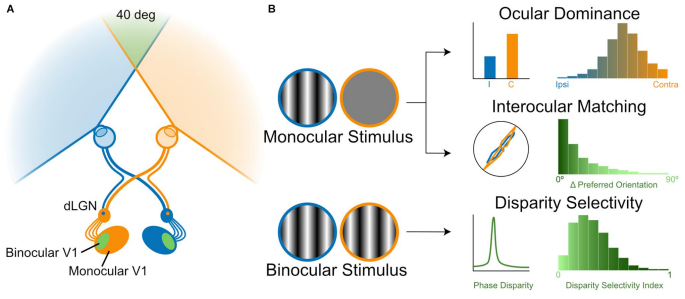
The objective is often an accurate, dense, metric representation of the observed geometry.

Human stereopsis, however, does not necessarily aim to produce a complete depth map. It supports perceptual judgments and actions, such as determining relative depth, identifying object boundaries, reaching toward objects, navigating, or maintaining binocular fixation. These objectives overlap, but they are not identical.

A human observer may understand that a chair is behind a table without reconstructing either object as a complete metric surface. This is a profound difference in what counts as a successful solution.

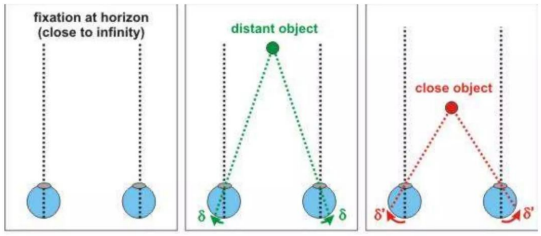
2.2 What biology actually contributes

Human stereopsis is based on several interacting mechanisms.



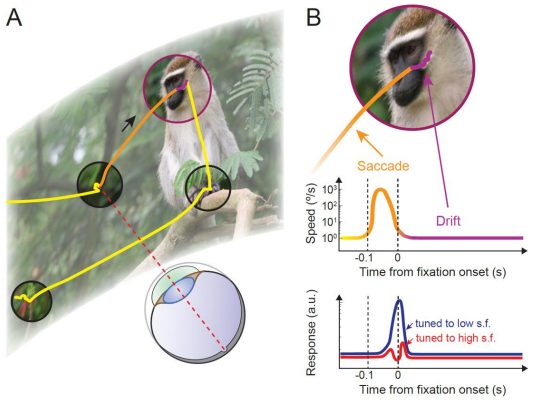
Neural disparity processing

Populations of binocular neurons are sensitive to disparities at different positions and scales. This provides evidence for distributed representations rather than a single explicit correspondence algorithm.



Vergence and fixation

The two eyes rotate to maintain binocular alignment around a selected target. This changes the distribution of retinal disparities and supports binocular fusion.



Sequential observation

Saccades and fixations provide different local measurements over time, while perceptual mechanisms maintain continuity across observations.

The classic review by Cumming and DeAngelis (2001) establishes an important distinction between early cortical disparity detectors and the additional processing needed for meaningful binocular correspondence and relative depth perception.

In this sense, biological stereo is not simply an alternative algorithm for computing disparity maps. It is embedded in a larger perceptual system.

2.3 Why these mechanisms were not extensively adopted

There are several reasons.

A. Biological mechanisms are not complete engineering algorithms.

We understand important components of neural disparity processing, but that does not give us a complete executable specification of human stereopsis. In particular, solving correspondence robustly across occlusions, ambiguous textures, and changing illumination remains difficult.

B. Biological vision prioritizes useful perception over complete metric reconstruction.

The human system is not universally accurate, nor does it recover metric depth equally well at every distance. Its solutions exploit context, experience, relative geometry, and perceptual assumptions.

C. Artificial systems have different constraints and opportunities.

A stereo camera with high-resolution sensors and powerful computing hardware can process an entire image simultaneously. A human-inspired sequence of fixations may introduce unnecessary latency for applications where dense measurements are already affordable.

D. Machine learning changed the nature of the competition.

Modern stereo models can learn strong correspondence features and scene priors directly from large datasets. For example, FoundationStereo (CVPR 2025) combines binocular matching with learned visual priors, while VGGT (CVPR 2025) predicts multiple geometric quantities from collections of images.

These methods demonstrate that there are powerful routes to 3D reconstruction without explicitly reproducing biological mechanisms.

2.4 But biological stereo has not actually been ignored

This is an important historical qualification.

Biological ideas have influenced computer vision through multiscale matching, disparity-selective filters, hierarchical processing, attention, vergence control, and active camera systems.

Researchers investigated robotic vergence and fixation-controlled stereo decades ago.

What has remained less central is the integration of these mechanisms into a unified, continuously operating perception system. That is where I see the more interesting gap.

2.5 An especially promising research question

Instead of asking:

Can we build a stereo matcher that works like the human visual cortex? I would ask:

Can active binocular control reduce the difficulty of stereo correspondence by changing the observations before attempting to solve the problem?

For example, suppose we have a binocular system with movable eyes and foveated sensing.

Rather than reconstructing the entire scene simultaneously, it could:

- Detect a candidate object at low resolution.
- Direct both foveae toward a selected region.
- Adjust vergence to bring the region into a favorable disparity range.
- Estimate local depth using geometric or learned correspondence.
- Extend acquisition toward uncertain surface regions and boundaries.
- Integrate successive observations into a persistent 3D representation.

The critical hypothesis is that a sequence of relatively simple, well-chosen local stereo problems may be more efficient or robust than one global, unconstrained problem.

This remains a hypothesis, not an established general result.

There is also an important geometric constraint: vergence does not magically create additional baseline or overcome the fundamental limits of triangulation. It can improve alignment, local sampling, and correspondence conditions, but the precision of metric depth remains constrained by viewing geometry and measurement noise.

2.6 A biological subtlety worth investigating

The conventional assumption that stereo precision is necessarily maximal at the fovea is more complicated than it initially appears.

A study published in Vision Research in December 2025 found that the location of best measured stereoacuity was not always foveal, even among normally sighted participants, under the tested stimulus conditions. This does not overturn the general importance of central vision, but it cautions against assuming that the optimal spatial sampling pattern for stereopsis is identical to that for ordinary visual acuity.

That observation reinforces a broader point:

A biologically inspired artificial retina should allocate resolution according to perceptual function, not simply copy the photoreceptor density distribution.

The optimal sampling pattern for recognizing an object may differ from the one for estimating depth, detecting motion, or identifying a geometric boundary. This could become a substantial research topic in its own right.

3. Does it make sense to teach a PhD-level course combining these topics?

Yes. In fact, I think the combination is stronger than either topic taught independently.

But I would make a distinction between a course about biological vision and a course that uses biological vision to rethink computational perception. For a PhD audience, I strongly favor the second. I can draft a syllabus for such a course.

3.1 A potential weakness of the course

There is a real danger that it becomes a broad but superficial mixture of neuroscience, geometry, robotics, and deep learning.

I would avoid that by organizing all the material around one computational question: How should a visual system acquire and integrate evidence?

Biological fidelity would be evaluated critically, not treated as inherently desirable.

Students would be encouraged to challenge biological inspiration when conventional methods perform better.

I would see this course as both a graduate teaching activity and a mechanism for generating research questions. In particular, it could connect students from mathematics, computing, robotics, graphics, and vision science who would otherwise approach these problems independently.

4. Are these promising scientific research directions worth pursuing?

Yes, but with different levels of scientific risk and different opportunities for original contributions.

I would distinguish the two directions.

Biologically Inspired Computer Vision is primarily a foundational research direction. Its central challenge is to extract computational principles from biology that improve artificial perception.

Active Foveated Reconstruction and Understanding of 3D Scenes is a more specific systems-research direction. It combines those principles with geometry, sensing, learning, and control to solve concrete problems.

The second may offer a clearer route to experimentally validated contributions, while the first provides the broader intellectual foundations.

4.1 Where is the actual research frontier?

This is not an unexplored field.

Active perception, next-best-view reconstruction, adaptive sensing, and learned stereo are already established research areas.

A 2025 review of next-best-view reconstruction surveyed more than 100 publications. Systems such as ActiveGAMER (CVPR 2025) already combine active exploration with 3D Gaussian Splatting and information-driven viewpoint selection.

Similarly, FOVI (ICML 2026) demonstrates how retina-inspired variable-resolution sampling can interface with contemporary deep vision models, reporting competitive performance with substantially fewer pixels and reduced computational requirements.

These results are encouraging, but they also establish a high standard for originality.

The opportunity is not simply to introduce foveation, active viewpoint selection, or learned stereo individually.

I see greater potential in developing a unified active visual system in which sensing, geometric measurement, semantic interpretation, and action are coupled through a coherent representation of uncertainty.

4.2 Five research questions I would prioritize

R1 — Does active foveation improve sensing efficiency?

Can nonuniform, dynamically controlled sampling attain comparable or better reconstruction and recognition quality than uniform sampling under equivalent acquisition and computational budgets?

Scientific contribution: measuring the value of selective sensing rather than merely reducing image resolution.

R2 — Can active binocular control simplify stereopsis?

Can fixation, vergence, and local multiscale processing improve the robustness of matching by restricting correspondence to favorable regions and disparity ranges?

Scientific contribution: establishing whether control genuinely simplifies correspondence after accounting for sensing and actuation costs.

R3 — How should binocular fixation interact with head motion?

Can an active observer maintain fixation while translating, using the resulting changing viewpoints to improve depth, resolve ambiguities, and reveal occluded surfaces?

Scientific contribution: integrating eye–head coordination with the geometry of information acquisition.

R4 — What is the appropriate 3D representation?

Should every visible region be reconstructed metrically, or can object-level geometry, uncertain depth, visual appearance, and semantic descriptions coexist in a unified scene representation?

Scientific contribution: connecting geometric observability with task-dependent scene understanding.

R5 — Can one coherent policy govern active scene understanding?

Can a controller decide when to inspect, measure, recognize, revisit, move, or stop without accumulating a large collection of task-specific exceptions?

Scientific contribution: a unified decision-making framework for sequential visual inference.

Of these, I consider R4 and R5 particularly important for turning a collection of interesting mechanisms into a general framework.

4.3 The near/far distinction may be more fundamental than it appears

This connects directly with our earlier discussion of foreground and background.

Consider the depth equation again:

$$Z = \frac{fb}{d}.$$

Assuming approximately constant uncertainty σ_d in measured disparity, first-order error propagation gives

$$\sigma_Z \approx \frac{Z^2}{fb}\sigma_d.$$

Thus, for a fixed stereo configuration, absolute depth uncertainty grows approximately quadratically with distance.

This observation has significant consequences.

A nearby tabletop object may require precise surface geometry for manipulation or contact. A distant building may need only approximate depth, a recognizable identity, and an appropriate appearance representation for a navigation task. The sky may be represented effectively as directional radiance without a finite metric surface.

Yet all of these regions contribute to scene understanding.

I would therefore investigate a scene representation in which the required geometric accuracy changes continuously with distance, uncertainty, and task.

A task-dependent hierarchy of scene representation

(a) Near field: Detailed surfaces, boundaries, metric depth, and object identity

(b) Intermediate field: Object layout, coarse geometry, relative depth, and semantics

(c) Far field: Directional appearance, semantic landmarks, depth ranges when useful



(a) (b) (c)

This is a proposed computational hierarchy, not a claim that human perception has three sharply separated processing stages. The appropriate representation depends on the task.

The interesting idea is not simply to separate the foreground and background into two layers.

It is to let the scene representation adapt to the observability and usefulness of each type of information.

I find the possibility of using a spherical, cyclopean scene representation especially appealing here. A common angular domain can support local object geometry, nonuniform visual evidence, semantic regions, and a far-field appearance model.

Such a representation could also integrate binocular observations and, eventually, head movement.

A 2026 review by Wilcox and Allison emphasizes that stereopsis and motion parallax share geometric foundations but have different measurement and perceptual constraints, reinforcing the importance of considering these cues jointly rather than independently.

4.4 A unifying mathematical formulation

I would organize the whole system around a belief state, rather than a conventional RGB-D image.

Let

$$B_t = P(S \mid O_{1:t}, A_{1:t}),$$

where S denotes the underlying scene, $O_{1:t}$ the accumulated visual observations, and $A_{1:t}$ the sensing actions already performed.

The belief state would include geometric estimates, object identities, confidence, and explicit distinctions between known and unknown information.

The next sensing action could then be chosen by

$$a_t^* = \operatorname{argmax}_a \left[\mathbb{E}[\Delta \mathcal{I}_{\text{task}} \mid B_t, a] - \lambda_s C_{\text{sensing}}(a) - \lambda_m C_{\text{motion}}(a) - \lambda_c C_{\text{compute}}(a) \right].$$

Here, $\Delta \mathcal{I}_{\text{task}}$ measures the expected improvement in information relevant to the current task.

The essential idea is simple:

Choose the next visual action according to what it is expected to contribute, taking its cost into account.

Importantly, this does not require a sophisticated reinforcement-learning controller from the beginning. A transparent, deterministic policy based on expected visibility, geometric uncertainty, and predicted information gain could be an excellent starting point.

4.5 The principal scientific difficulties

I would consider the following challenges potentially decisive.

Challenge	Why it is hard
Nonuniform binocular correspondence	The two retinas have gaze-dependent sampling patterns that may not align naturally
Geometric observability	Some ambiguities cannot be resolved without changing viewpoint or baseline
Integration across fixations	Measurements must remain consistent despite gaze, calibration, and pose changes
Reliable uncertainty	The controller must distinguish unknown geometry from failed measurement and unsupported inference
Active control and stopping	A policy must avoid redundant observations, oscillations, and unjustified termination
Biological relevance	A biologically inspired mechanism must be demonstrated to improve some measurable objective
Real-world generalization	Noise, dynamic scenes, illumination, and calibration errors may invalidate simplified assumptions

I would emphasize one crucial limitation of a fixed-head architecture.

Eye rotations and saccades do not generally create the viewpoint diversity needed to reveal genuinely occluded surfaces.

They can select new image regions and improve sampling, but rotating an ideal camera around a fixed optical center does not create translational parallax.

Consequently, the progression from fixed-head foveation to controlled head translation is not merely an engineering extension. It changes which geometric information can actually be obtained.

4.6 How I would establish scientific value

I would insist on comparisons under equivalent constraints.

For example, compare four acquisition strategies:

1. Uniform, passive stereo.
2. Fixed foveation with passive acquisition.
3. Active foveation with gaze control.
4. Active foveation with controlled eye–head movement.

Hold the stereo inference method and other relevant components fixed when testing the acquisition strategy.

Evaluate depth accuracy, surface completeness, boundary quality, semantic accuracy, acquisition bandwidth, motion costs, computation, and stopping behavior.

Ground-truth segmentation and depth from Blender would be perfectly reasonable for controlled diagnostic experiments. But experiments claiming autonomous perception must eventually remove those privileged inputs from the inference and decision-making loop.

The October 2026 review Active Visual Perception: Opportunities and Challenges explicitly highlights the importance of matched sensing budgets, component-level comparisons, and measuring the information gained immediately after a sensing action—not merely final task performance.

That is exactly the methodological discipline this research needs.

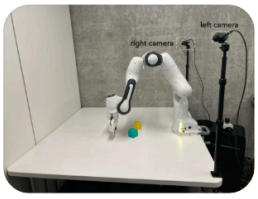
5. Can these ideas lead to practical applications and innovations?

Yes, particularly when sensing, computing, communication, energy, or movement are constrained.

The commercial opportunity is less obvious for a conventional offline 3D reconstruction pipeline that already has hundreds of high-quality images.

It becomes more compelling for intelligent systems operating continuously, especially when they cannot afford to process every observation at maximum resolution.

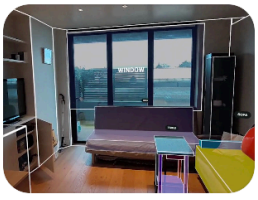
5.1 Promising applications



Robotics and manipulation

An intelligent binocular sensor could focus attention on candidate objects, recover detailed local geometry needed for grasping, and move to resolve uncertainty.

Principles: active stereo, object-centric geometry, uncertainty-driven acquisition.



Augmented and mixed reality

A headset could maintain detailed 3D geometry near interaction regions while representing distant surfaces using lower-cost appearance and approximate geometry.

Principles: multiresolution perception, task-dependent geometry, predictive gaze and head tracking.



Industrial inspection and autonomous exploration

A visual system could identify potentially important components, inspect them at high resolution, and choose new viewpoints where its geometric or semantic understanding is insufficient.

Principles: information gain, active inspection, selective sensing, semantic attention.



Immersive 360° media and telepresence

A system could maintain a panoramic visual world and selectively acquire accurate 3D geometry around objects that the observer wants to examine or interact with.

Principles: spherical visual representation, selective depth acquisition, persistent scene memory.



Energy-efficient and event-based vision

Combining spatially adaptive acquisition with asynchronous sensing could support responsive visual systems where bandwidth and power are limited.

Principles: sparse encoding, adaptive temporal resolution, event-driven computation.

For the last application, there is already encouraging enabling technology. Event-based stereo is an active research field, and electronically reconfigurable multifoveated vision sensors have been demonstrated. The integration of these capabilities into a complete active 3D perception system remains an interesting engineering challenge.

5.2 Three innovations I find particularly attractive

Innovation A — An active binocular visual sensor

Imagine a compact binocular head with a wide contextual field and two high-resolution regions whose viewing directions can be controlled.

The system would use peripheral sensing to detect where important information may be found, then direct local high-resolution acquisition toward those regions.

Its output would not simply be stereo images or an instantaneous depth map.

It would maintain a continuously updated 3D scene estimate with explicit geometric and semantic confidence.

This is potentially useful for robotics, inspection, and lightweight autonomous systems.

The fundamental innovation is controlling the acquisition of visual evidence rather than treating sensing as a fixed input process.

Innovation B — A perceptually organized 3D world representation

This would combine several representations in one system:

- Detailed 3D geometry for nearby and important objects.
- Approximate geometry and semantics for intermediate regions.
- Directional appearance and coarse depth information for distant backgrounds.
- Explicit uncertainty and observation history across all regions.

The goal would be a world model that is neither a dense geometric reconstruction everywhere nor a purely semantic scene description.

It would retain whatever information is appropriate for interaction, visual continuity, and subsequent perception.

This is especially attractive for immersive environments, virtual reality, mixed reality, and panoramic scene understanding.

It is important to distinguish it from conventional foveated rendering, which concentrates rendering resources where the user looks.

Foveated rendering is already well developed, including in modern neural rendering systems. The additional innovation proposed here is to apply adaptive allocation to the acquisition and reconstruction of the world itself.

Innovation C — A scene-understanding system that knows what to look at next

This is perhaps the most ambitious and scientifically interesting innovation.

Imagine a robot observing an unfamiliar environment. Instead of attempting to reconstruct everything immediately, it maintains an evolving representation of the scene.

It distinguishes what it has observed, what it has measured reliably, what it has inferred, and what remains unknown. It might ask:

Is this surface boundary real? Is that dark region an opening or a shadow? Is this object partly occluded? Do I need another viewpoint to distinguish these possibilities?

The system would choose its next observation accordingly.

This could combine learned recognition models with geometric reasoning and active control.

The underlying principle is perception as sequential hypothesis testing through action.

A valuable consequence would be a more auditable form of visual intelligence: the system could explain which conclusions are supported by measurements and which remain predictions.

6. Bringing the two research directions together

Although your questions were organized into five parts, I think they point toward one additional issue: how to combine the ideas without building an unnecessarily complicated system.

I would propose a progression from a scientific prototype to a more general perceptual architecture.

Proposed research progression

1. Active foveated stereo with a fixed head

   Establish the effects of retinal sampling, gaze selection, vergence, and sequential integration. Use simple static scenes and a controlled observation budget.

   Core question: Is purposeful acquisition better than passive sampling when the geometric viewpoint is fixed?

2. Object-centric 3D scene reconstruction

   Extend the system from isolated surface exploration to multiple objects, explicit boundaries, persistent geometry, and semantic labels.

   Core question: Can one coherent acquisition policy support object discovery, growth, verification, and stopping?

3. Binocular fixation combined with controlled head motion

   Introduce small translations and study the interaction between binocular disparity, motion parallax, fixation stability, and previously occluded geometry.

   Core question: Which movements genuinely improve geometric observability?

4. Task-dependent scene representation

   Integrate near-field geometry, middle-distance object organization, and far-field appearance with explicit uncertainty.

   Core question: How much geometry does a system need to understand and use a scene?

5. Learning-enhanced active perception

   Replace selected hand-designed components with learned predictors while keeping the measurement geometry, uncertainty accounting, and experimental comparisons explicit.

   Core question: Can learning improve the active sensing strategy without losing geometric consistency and interpretability?

This progression fits naturally with the Tabletop and Classroom scenes we've discussed, as well as the existing approach of using a simulation environment to separate scientific questions from the complications of physical hardware.

It also offers a useful division of responsibilities between geometry and machine learning.

Geometry determines what can be measured and how observations constrain the scene. Learning helps deal with ambiguity, appearance variability, object semantics, and decisions that are difficult to encode analytically.

Neither should be expected to replace the other.

7. My overall assessment

I would summarize the scientific and technological prospects as follows.

Dimension	Assessment	Reason
Scientific foundations	Very strong	Connects established geometry, neuroscience, information theory, and control
PhD education	Very strong	Encourages cross-disciplinary reasoning around a concrete mathematical problem
Research originality	Promising but demanding	Individual components are established; integration alone is insufficient for novelty
Near-term demonstrability	Strong	Controlled simulations can test key hypotheses without specialized hardware
General-purpose scene understanding	High potential, high risk	Robust control, observability, and memory integration remain difficult
Practical applications	Promising in selected domains	Most attractive under sensing, computation, motion, or energy constraints

My strongest conviction is that active visual understanding should not be judged by whether it reconstructs the entire scene more accurately than a conventional system with unlimited images and computation.

That would miss much of the point.

It should be judged by whether it can achieve useful understanding with less information, less computation, fewer observations, and better control over uncertainty.

At the same time, we should not assume that biology supplies an optimal solution. An artificial visual system may benefit from biological principles while using sensors, representations, and computational strategies that evolution never produced.

One final conceptual question

There is a distinction that I think deserves to become central to this research:

Is the objective of visual intelligence to reconstruct the world, or to acquire and maintain sufficient knowledge of the world to act intelligently?

The first leads naturally to dense geometry, stereo matching, and 3D reconstruction.

The second leads to attention, active sensing, adaptive resolution, semantic understanding, memory, and decision-making.

The two are complementary, but the second is the more general formulation.

My preference would therefore be to organize the research around a single long-term goal:

> *Develop a computational theory and experimental framework for active, foveated, binocular scene understanding, in which a visual system selects and integrates observations to build task-appropriate geometric and semantic representations of its environment.*

That is broad enough to sustain a PhD-level course and a multi-year research program, yet specific enough to support a sequence of falsifiable experiments.

And I think its most important intellectual contribution could be a change in what we consider the fundamental unit of computer vision: not the image, not even the reconstructed scene, but the evolving relationship between an observing system and the world it seeks to understand.
