Gaze-Variance-Adaptive Foveation for VR Cloud Gaming: Dynamically Adjusting Foveal Region and Compression Falloff Based on Recent Gaze Activity

Sam Jett, Rohan Parekh

Abstract:

Virtual reality (VR) technology is increasingly used for gaming and other immersive applications. However, delivering high visual fidelity in a headset form-factor remains challenging. VR is trapped in a space with several competing factors including performance, quality, comfort, and mobility. It is impossible to have a "super VR" headset that can run at high framerates and realistic resolutions while being light, portable, affordable, and wireless; high-end desktop performance cannot fit comfortably onto a headset. One attempt at resolving this is to stream wirelessly from a server instead of over a cable. This usually is crippled by high latency and poor bitrates. EyeNexus is a research system that tackles this problem by combining foveated (gaze-driven) spatial compression (FSC) with foveated video encoding (FVE). It uses real-time eye-tracking features of certain headsets to concentrate high bitrate where the user is looking and adapts foveation to network conditions. The result of this is reduced latency and improved perceived quality for wireless VR streaming from a remote game server. We reproduce the EyeNexus experiment to validate these results, and we extend the EyeNexus eye-tracking methods to measure gaze variance over time and make changes to the FSC. We evaluate the ability of our extensions to save additional bandwidth or improve quality, depending on network conditions, and we evaluate if our methods improve the latency of the FVE.

1. Introduction

The virtual reality (VR) market is expanding rapidly, estimated to be worth 259 billion by 2034 [1]. Some VR headsets have evolved to support wireless streaming, allowing greater freedom of movement by removing bulky connections. Cloud gaming further enhances accessibility by offloading computation to remote servers, eliminating the need for expensive local gaming hardware. However, wireless networks are at odds with the user's desire for high bitrates and low latency. The datarates for VR are massive: uncompressed 5.7K 360° video (the highest supported by 360° cameras) would require streaming at ~7Gbit/s. Yet, this only achieves approximately 27% of the 60 samples-per-degree standard considered necessary for retina quality [2].

To ease these challenges, approaches like Google Congestion Control (GCC) [3] and open-source platforms like Air Light VR (ALVR) [4] employ adaptive bitrate (ABR) streaming, though they often sacrifice visual quality to preserve low latency. We've also seen foveated techniques that exploit the non-uniform acuity of human sight, which resolves fine detail only within a narrow region (around 1.5-5° of visual angle) while peripheral vision is coarser [5]. By concentrating quality around a viewer's gaze point, foveated video encoding (FVE) can reduce bandwidth demands with less perceptible quality loss.

Recent work has made strides in combining these approaches. EyeNexus [6] introduces a novel system that integrates real-time foveated spatial compression (FSC) with foveated video encoding (FVE), using a Gaussian-distributed foveation model that dynamically adjusts the foveation region based on both real-time gaze data and available bandwidth. EyeNexus demonstrates latency reductions of up to 70.9% and perceptual visual quality improvements of up to 24.6% compared to state-of-the-art baselines.

In this paper, we reproduce the results of EyeNexus and extend it by incorporating recent gaze variance into the foveation model. Our hypothesis was that adjusting the foveal region could allow us to save on bitrate during times of low variance (low amounts of looking around) and improve user perception during times of high variance. We evaluate the reproduction and extension against the baseline by recording latency, bitrate, and the CDF of latencies for both methods.

2. Background and Related Work

2.1 Human Visual Perception and Foveation

Humans have a field of view of approximately 220° horizontally by 135° vertically [9], yet only a small central region—the fovea, spanning roughly 1.5° to 5° of visual angle—is capable of resolving fine spatial detail at up to 60 cycles per degree [10]. Outside the fovea, visual acuity drops sharply [5, 11]. These perceptual characteristics form the basis for foveated graphics techniques that degrade image quality of regions in the viewer's periphery to reduce computational load and bandwidth usage without perceptible quality loss to the user.

Importantly, the human visual system does not maintain constant sensitivity. During saccadic eye movements, saccadic suppression temporarily reduces visual perception for a period of 50 to 200 ms [7, 12]. These temporal dynamics of gaze behavior suggest that a foveation system could adapt to both where the user is looking and also to how their gaze is behaving temporally, which was the motivation for our project.

2.2 Foveated Video Encoding and Spatial Compression

ALVR [4] is an open-source VR gaming platform that streams from a gaming PC to a VR headset over Wi-Fi, adjusting the target bitrate based on frame interval and throughput statistics maintained in a sliding window. It applies uniform encoding across the entire frame, treating all regions as equally important regardless of where the user is looking. This means the bitrate is spread thin instead of being focused on what the user is at.

Foveated video encoding (FVE) exploits the non-uniform nature of human perception by assigning variable quantization across a video frame, with less quantization near the gaze point and more quantization in the periphery [13]. Foveated spatial compression takes a complementary approach by reducing the pixel density in peripheral regions through 2D spatial warping before encoding, reducing bandwidth usage [14]. However, these approaches do not adapt foveation to network conditions.

2.3 The EyeNexus System

EyeNexus [6] addresses the limitations of prior work by introducing a system that combines real-time gaze-driven FSC with adaptive gaze-driven FVE. The system maps VR gaze coordinates to screen space, applies dynamic FSC to reduce frame resolution while preserving foveal detail, and then encodes the compressed frame using a Gaussian-distributed quality assignment centered on the gaze point. A foveation controller governs the size of the high-quality foveal region with an AIMD congestion control scheme, expanding the foveal region when bandwidth is plentiful and shrinking it during congestion.

In their evaluations, EyeNexus achieves a superior mean motion-to-photon (MTP) latency (69 ms) and perceptual visual quality among all baselines. However, its foveation model adapts only to network conditions and the instantaneous gaze position. It does not consider temporal patterns of gaze behavior. Our work extends EyeNexus by incorporating recent gaze variance as an additional signal for modulating the foveation parameters, enabling more fine-grained adaptation to the user's perceptual state.

2.4 Gaze Dynamics and Perceptual Sensitivity

During saccades (rapid movements of the eye when adjusting the point of focus), sensitivity to fine detail is reduced [7]. Research demonstrates that viewers frequently fail to detect even large changes in visual scenes when those changes coincide with such disruptions [15], suggesting that moments of high gaze variance can be leveraged to normalize visual quality across the screen without affecting user perception.

To our knowledge, no prior work in foveated video encoding has incorporated real-time gaze variance as a parameter for dynamically adjusting the foveation model. Our work bridges this gap by proposing a gaze-variance-adaptive foveation strategy that modulates both the foveal region radius and compression falloff steepness in response to the user's recent gaze behavior.

3. Implementation

To reproduce the results, we created the simplest setup that we could with the hardware we had. We used a PC with an AMD 5600X CPU (6 cores, 12 threads), an SSD, and an RTX 3090 GPU. The headset was a Meta Quest Pro. Connecting them is a Netgear 6400v2 router, with 1GbE to the server and 5GHz wireless to the headset (peak ~300Mb/s). Nothing else was on the network, and there was minimal wireless traffic in the vicinity of the experiment.

EyeNexus defines this Gaussian profile for the foveated region (QO is defined by the EyeNexus authors as an offset from the quantization parameters):

$$QO(i,j) = QO_{max} - QO_{max} \times \exp\left(-\frac{(Distance(i,j))^2}{2C^2}\right)$$

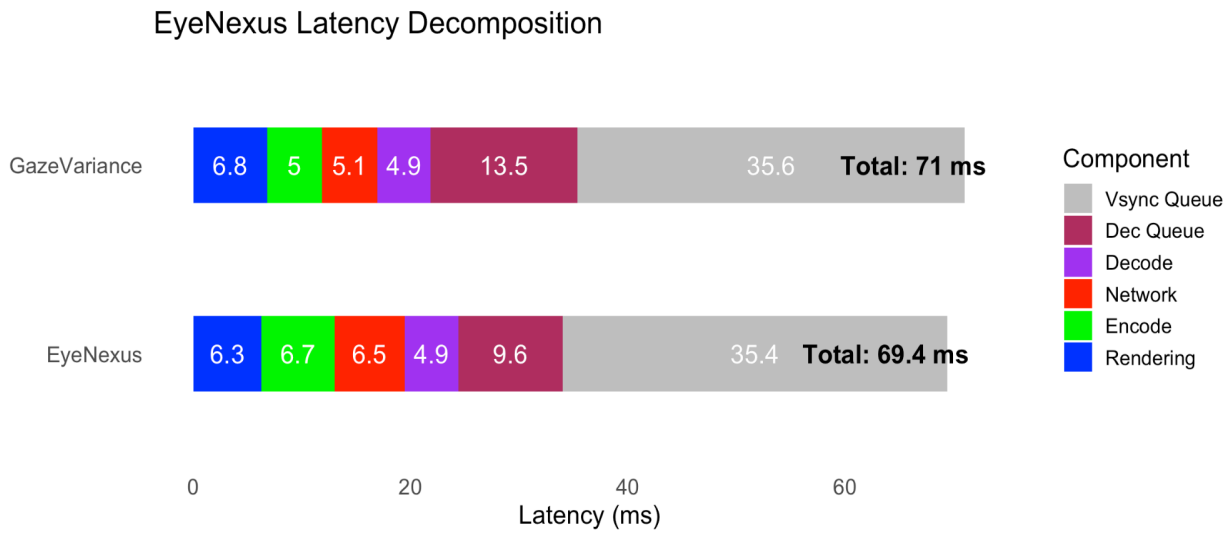
In the base EyeNexus, the C (sigma) is defined exclusively by the network controller. In our gaze variance implementation, we have an effective C:

```python
def clamp(x, lo, hi):
    if x < lo then return lo
    if x > hi then return hi
    return x

C_MIN = 2.0
C_MAX = 80.0
w_f = 0.15   // fixation confidence weight

def compute_C_eff_fixation(C1, f):
    if f is None or f < 0 or f > 1:
        return clamp(C1, C_MIN, C_MAX)
    scale = 1.0 - w_f * f
    return clamp(C1 * scale, C_MIN, C_MAX)

def QO(i, j, X_QP, Y_QP, C, QO_max):
    dist_sq = (i - X_QP)^2 + (j - Y_QP)^2
    distance = sqrt(dist_sq)
    exponent = -dist_sq / (2 * C^2)
    return QO_max - QO_max * exp(exponent)
```

Fixation confidence is a measure (from 0 to 1) of how reliably the user is in a stable fixation. High confidence justifies a tighter high-quality foveal region (smaller effective C) and stronger peripheral compression; low confidence leads to a wider high-quality region to avoid visible artifacts when gaze is moving or uncertain. See the second function in the above code block; higher confidence shrinks the spread. The confidence is approximated by $(1 + \sigma^2)^{-1}$. Reducing reliance on fine detail during non-fixation and use of gaze stability as a proxy for fixation are supported by prior research [16, 17].

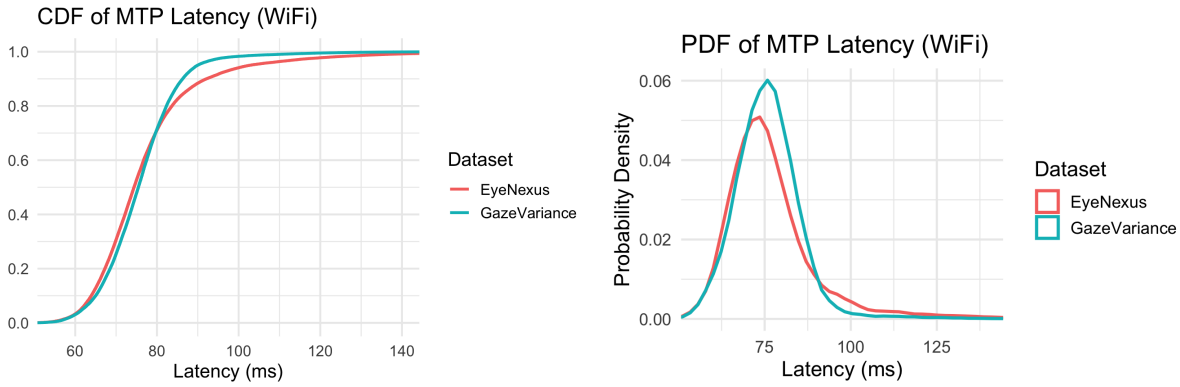
4. Evaluation

4.1 End-to-End Latency



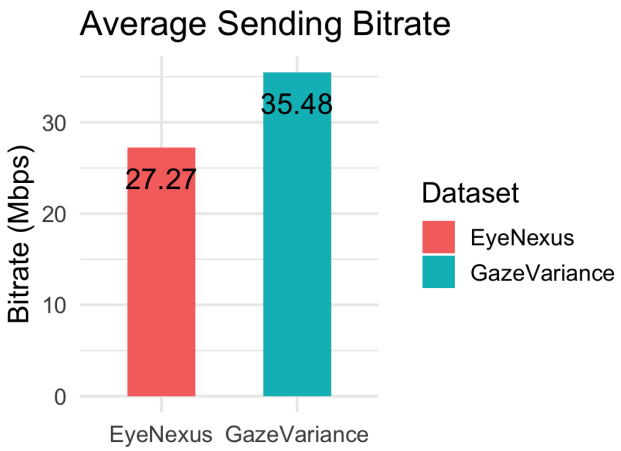
Our implementation achieves a mean end-to-end latency of 71.0 ms compared to 69.4 ms for baseline EyeNexus (our reproduced value is equal to what was reported in the original paper). This 1.6 ms overhead remains substantially lower than other baselines reported by Wu et al. [6]: GCC (77.7 ms), ALVR (78 ms), FovOptix (79.3 ms), and CGFVE (80.9 ms).

The latency decomposition reveals where the difference originates. Rendering latency increases slightly due to the gaze variance computation. However, encoding latency and network latency both decrease. During high-variance gaze periods, our model scales down the effective foveal region, producing more aggressive compression and frames that encode and transmit faster. The decode queue latency increases, likely because our higher average sending bitrate (Section 4.2) produces larger frames that consume more buffer space on the resource-constrained Quest Pro.



The CDF and PDF of MTP latency show that our implementation has comparable or slightly tighter tail behavior than baseline EyeNexus, with no evidence of latency spikes or instability. This suggests the gaze-variance model may help smooth worst-case latency by proactively tightening foveation during rapid gaze movement, preventing the demand spikes that occur when a large foveal region tracks a fast-moving gaze point.

4.2 Sending Bitrate



GazeVariance achieves an average sending bitrate of 35.48 Mbps compared to 27.27 Mbps for baseline EyeNexus, a 30% increase.

This higher average arises because the network controller (AIMD) treats the savings during fixation as spare capacity and additively increases C, so that during saccades and gaze movement C is larger than baseline and the (non-normalized) Gaussian allocates more bits to a wider high-quality region.

This additional bandwidth is not wasted. It is allocated specifically during periods when the user's gaze is unpredictable and may fixate anywhere in the scene, ensuring that wherever the gaze lands, image quality is higher than it would be under baseline EyeNexus's narrower foveal region. During fixation, when the user is attending to a single point, the system saves bandwidth by tightening the foveal region. This tradeoff is justified by the perceptual literature on saccadic suppression [7] and change blindness [15], which shows that peripheral quality changes during stable fixation are among the hardest artifacts for users to detect.

4.3 Visual Quality

We were unable to reproduce the EWSSIM and EWPSNR measurements from the original paper. These visual quality metrics require comparing delivered frames against uncompressed reference frames using synchronized gaze coordinates. Wu et al. neither describe the capture pipeline for these measurements nor include instrumentation in the codebase for intercepting reference frames, synchronizing frames, or aligning gaze data. Implementing this from scratch was beyond our project's scope.

Despite this, the combination of our latency and bitrate results provides strong evidence for improved perceived quality. We send 30% more data while maintaining comparable latency, with the additional bits directed toward frames where gaze behavior is most unpredictable. By maintaining higher quality across a broader region during scanning, we ensure that the image quality is already high wherever the gaze settles.

4.4 Reproduction Notes

Our reproduction revealed that the MTP latency reported in the original paper excludes two logged columns: game_latency_ms (game simulation time) and composite_latency_ms (left/right eye image compositing time). Both are fixed overheads that do not vary across streaming protocols, so their exclusion does not affect comparative results. However, the reported MTP latency is the streaming pipeline latency rather than true motion-to-photon time. We adopt the same convention for comparability.

5. AI Use

First, given the sheer size of the EyeNexus codebase (and the ALVR code it is built on top of), we had AI break down and search the structure of the codebase to help us determine what sections we would need to modify to add gaze-variance foveation. Additionally, we used AI to keep the documentation up-to-date with our code changes. Finally, we used AI to assist in background research and finding relevant work.

References

[1] Fortune Business Insights, “Virtual Reality (VR) in Gaming Market Size, Share & Industry Analysis,” 2024.

[2] L. Hsiao, B. Krajancich, P. Levis, G. Wetzstein, and K. Winstein, “Towards Retina-Quality VR Video Streaming: 15 ms Could Save You 80% of Your Bandwidth,” ACM SIGCOMM Computer Communication Review, vol. 52, no. 1, Jan. 2022.

[3] G. Carlucci, L. De Cicco, S. Holmer, and S. Mascolo, “Analysis and Design of the Google Congestion Control for Web Real-Time Communication (WebRTC),” in Proc. MMSys ’16, 2016.

[4] ALVR, “Air Light VR,” 2024. https://github.com/alvr-org/ALVR

[5] H. Strasburger, I. Rentschler, and M. Jüttner, “Peripheral Vision and Pattern Recognition: A Review,” Journal of Vision, vol. 11, no. 5, 2011.

[6] Z. Wu, A. Alhilal, Y. H. Tsui, M. Siekkinen, and P. Hui, “EyeNexus: Adaptive Gaze-Driven Quality and Bitrate Streaming for Seamless VR Cloud Gaming Experiences,” Proc. ACM Netw., vol. 3, CoNEXT4, Article 42, Dec. 2025.

[7] J. Ross, M. C. Morrone, M. E. Goldberg, and D. C. Burr, “Changes in Visual Perception at the Time of Saccades,” Trends in Neurosciences, vol. 24, no. 2, pp. 113–121, 2001.

[8] M. Rucci and M. Poletti, “Control and Functions of Fixational Eye Movements,” Annual Review of Vision Science, vol. 1, pp. 499–518, 2015.

[9] A. Knapp, “An Introduction to Clinical Perimetry,” Archives of Ophthalmology, vol. 20, no. 6, 1938.

[10] M. F. Deering, “The Limits of Human Vision,” in Proc. 2nd International Immersive Projection Technology Workshop, 1998.

[11] L. N. Thibos, F. E. Cheney, and D. J. Walsh, “Retinal Limits to the Detection and Resolution of Gratings,” J. Opt. Soc. Am. A, vol. 4, no. 8, pp. 1524–1529, 1987.

[12] R. H. S. Carpenter, Movements of the Eyes, 2nd Rev. Pion Limited, 1988.

[13] G. K. Illahi, T. Van Gemert, M. Siekkinen, E. Masala, A. Oulasvirta, and A. Ylä-Jääski, “Cloud

Gaming with Foveated Video Encoding," ACM Trans. Multimedia Comput. Commun. Appl., vol. 16, no. 1, Article 7, Feb. 2020.

[14] T. Kämäräinen and M. Siekkinen, "Foveated Spatial Compression for Remote Rendered Virtual Reality," in Proc. MetaSys '23, 2023.

[15] R. Rosenholtz, "Capabilities and Limitations of Peripheral Vision," Annual Review of Vision Science, vol. 2, pp. 437–457, 2016.

[16] R. Uramune, S. Ikeda, H. Ishizuka, and O. Oshiro, "Fixation-based Self-calibration for Eye Tracking in VR Headsets," arXiv:2311.00391, 2024.

[17] S. Idrees, M. P. Baumann, F. Franke, T. A. Münch, and Z. M. Hafed, "Perceptual Saccadic Suppression Starts in the Retina," Nature Communications, vol. 11, no. 1, p. 1977, 2020.