

iSR: Super-resolution for Immersive Cloud VR Gaming Platforms

Ghazaleh Bakhtiariadzad
gba50@sfu.ca
Simon Fraser University
Burnaby, BC, Canada

Haseeb Ur Rehman
hurre100@uottawa.ca
University of Ottawa
Ottawa, ON, Canada

Shervin Shirmohammadi
shervin.s@uottawa.ca
University of Ottawa
Ottawa, ON, Canada

Ihab Amer
iamer@aus.edu
American University of Sharjah
Sharjah, United Arab Emirates

Mohamed Hefeeda
mhefeeda@sfu.ca
Simon Fraser University
Burnaby, BC, Canada

Abstract

Cloud-based Virtual Reality (VR) gaming enables immersive experiences without the need for costly high-end consumer hardware. However, it imposes substantial bandwidth requirements due to the need to stream high-resolution, high frame rate, and stereoscopic frames to maintain immersion and prevent motion sickness. Existing techniques like foveated rendering and encoding face challenges such as reliance on costly eye-tracking hardware and sensitivity to sudden gaze shifts. In addition, prior super-resolution methods can improve fidelity but are often too computationally heavy for practical deployment. To address these challenges, we propose iSR, a system that integrates stereo-aware colorization and super-resolution to reduce transmission cost while preserving visual quality. The key idea of iSR is that it first downsamples both stereo views to reduce the total number of transmitted pixels. Then, it transmits one view in full color and the other in monochrome. This removes redundant chrominance information and further reduces the required bandwidth. On the client side, iSR reconstructs full-color, high-resolution stereo frames by transferring chroma between views and enhancing spatial resolution. Extensive experiments across multiple VR games show that iSR achieves substantial bitrate reductions while maintaining high visual fidelity. These results highlight its potential for enabling high-quality VR streaming in bandwidth-limited environments.

CCS Concepts

• Information systems → Multimedia streaming.

Keywords

Cloud VR gaming, bandwidth optimization, super-resolution, stereo-aware colorization

ACM Reference Format:

Ghazaleh Bakhtiariadzad, Haseeb Ur Rehman, Shervin Shirmohammadi, Ihab Amer, and Mohamed Hefeeda. 2026. iSR: Super-resolution for Immersive Cloud VR Gaming Platforms. In *ACM Multimedia Systems Conference 2026 (MMSys '26)*, April 04–08, 2026, Hong Kong, Hong Kong. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3793853.3795746>



This work is licensed under a Creative Commons Attribution 4.0 International License. *MMSys '26, Hong Kong, Hong Kong*
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2481-7/2026/04
<https://doi.org/10.1145/3793853.3795746>

1 Introduction

Virtual Reality (VR) gaming offers highly immersive 3D experiences by rendering complex scenes at high resolution and high frame rates for each eye, creating a strong sense of presence [34]. This realism has fueled rapid adoption: the global VR gaming market was valued at 32.5B USD in 2024 and is projected to reach 109.6B USD by 2030 [12]. However, delivering such experiences requires substantial computing resources to maintain the fidelity and latency needed for immersive VR. State-of-the-art (SOTA) headsets already exceed 4K resolution and operate at 90–144 frames per second (fps) [41]. Moreover, motion-to-photon latency should remain below 20 ms to prevent discomfort or motion sickness [38]. These requirements make VR among the most resource-intensive interactive media.

Cloud computing offers a promising solution for VR gaming by offloading the complex rendering operation to powerful remote servers and streaming the resulting frames to lightweight headsets [44]. However, this shifts the challenge from expensive, high-end local hardware to network capacity. This is because VR gaming requires streaming *stereoscopic* frames at high resolutions and high frame rates, which requires substantial network bandwidth.

We conducted a measurement study to approximately assess the bandwidth requirements of VR cloud gaming platforms. Specifically, we subscribed to the Shadow PC VR cloud gaming platform and measured the bandwidth consumption of multiple VR gaming sessions using Wireshark [37]. The setup of our experiment is illustrated in Figure 1. The cloud servers were located in Portland, Oregon, while the player was located in a North America-based lab. The player used a Meta Quest 3 headset connected via a 5 GHz Wi-Fi network with approximately 500 Mbps bandwidth. We recorded three one-minute sessions for each of three VR games (Beat Saber, Pavlov VR, and VRChat), streamed at 4K stereo resolution. We then computed per-game distributions of instantaneous bitrate across these sessions. The resulting bandwidth usage distributions are shown in Figure 2. As the figure illustrates, all three games exhibit median bitrates between 72 and 78 Mbps, which demonstrates the high bandwidth challenge facing cloud VR gaming platforms. Our results are in line with prior experiments reported in [6, 27].

Recent approaches aim to alleviate the bandwidth burden through foveated rendering and encoding, which leverage the non-uniform acuity of the human visual system by prioritizing quality in the foveal region [2]. To operate reliably, these methods require precise eye-tracking and low-latency gaze prediction, which add hardware

cost [9] and may introduce artifacts during rapid gaze shifts common in VR games [29].

A potential approach to reduce bandwidth is to utilize super-resolution (SR), which transmits low-resolution versions that are later upscaled at the client. Several works have proposed SR methods for various multimedia systems, including 2D video streaming [20], stereoscopic video streaming [39], and omnidirectional/360° video streaming [8, 19, 40]. These works, however, are not suitable for *interactive* VR cloud gaming systems, as they introduce substantial computational costs and, more critically, they have long processing latencies that VR cloud gaming systems cannot tolerate. To demonstrate this limitation, we evaluated the SOTA SR model for VR content, which is CVHSSR [47]. Our experiments show, for example, that the inference time for this model to upscale the two stereo views by a factor of $4\times$ is approximately *3 seconds* on a decent workstation. This latency is at least an order of magnitude longer than that of realistic VR cloud gaming systems, limiting the potential deployment of this model in practice.

In addition, common SR models from industry, such as NVIDIA DLSS [28], AMD FSR [3], and Intel XeSS [18], are designed for *local* gaming consoles and cannot be used for cloud VR gaming platforms. This is because all such SR methods require access to the game renderer's data structures [33], which are only available on the game server in the cloud, not at the client.

In this paper, we introduce iSR (immersive Super Resolution), a system designed to reduce the bandwidth requirements for interactive VR cloud gaming while preserving visual quality. iSR exploits the redundancy between stereo views by transmitting one downsampled color view and one downsampled monochrome view. It then reconstructs full-color, high-resolution stereo views on the client by integrating colorization and super-resolution components. This design eliminates redundant chrominance transmission and ensures stereo consistency.

The contributions of this paper can be summarized as follows:

- We present iSR, a bandwidth-saving system that integrates stereo-aware colorization and super-resolution to reduce transmission cost while maintaining perceptual quality. Unlike existing SR methods, iSR does not require accessing game source code or gaze tracking hardware, making it more deployable in existing VR streaming platforms without system-level modifications.
- We collect and curate a high-resolution VR gaming dataset from popular games to simulate realistic streaming scenarios. **We will share this dataset with the research community.**
- We implement iSR and conduct extensive experiments to evaluate its performance. The results demonstrate that iSR can upscale VR gaming streams with different scaling factors while achieving high quality (34+ dB) and substantially reducing network bandwidth. iSR reduces the bandwidth by up to 56% while running in real time.

2 Background and Related Work

Virtual Reality. VR systems provide immersion by delivering separate high-resolution images to each eye. Depth perception arises from stereopsis: two forward-facing eyes, separated by ~ 6.3 cm, perceive slightly shifted views that the brain fuses into a coherent 3D scene [13]. Modern Head-Mounted Displays (HMDs) replicate

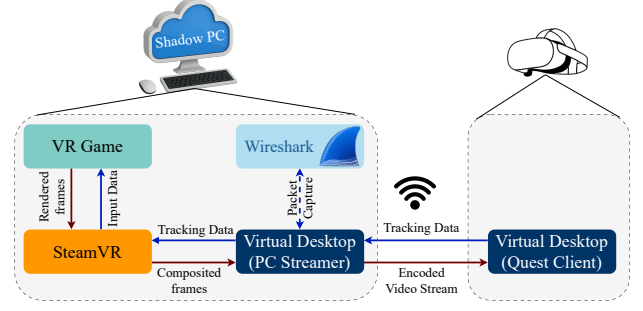


Figure 1: Experimental setup for measuring bandwidth usage in cloud VR gaming.

this process with dual displays and head-tracking sensors that update images with every movement. To maintain immersion and prevent discomfort, these displays must operate at very high fidelity, 4K-8K per eye at 90–144 Hz, with motion-to-photon latency under 20 ms [7]. These requirements make VR one of the most resource-intensive applications in modern computing. A single stereo frame at 4K per eye corresponds to ~ 16 million pixels, and sustaining 90 Hz stereo playback requires over 2.8 billion pixels per second. This enormous rendering and transmission load is a key bottleneck for widespread VR adoption.

VR Cloud Gaming. Cloud computing addresses the hardware cost barrier of VR by rendering on powerful servers and streaming the results to lightweight HMDs [44]. However, this shifts the challenge from computation to bandwidth and latency.

VR streams carry two high-resolution views instead of one, which roughly doubles the transmitting data compared to conventional 2D streams. Even after compression, a 4K stereo VR stream typically requires more than 70 Mbps as indicated in [6, 27] and demonstrated by our experiments in §1. At higher resolutions (e.g., 8K per eye), the bitrate becomes impractical for most players.

VR applications demand ultra-low motion-to-photon latency (~ 20 ms). This budget must account for server rendering, encoding, network round-trip, decoding, and display refresh. Unlike conventional video streaming, even small delays break immersion and may cause motion sickness [14].

Stereo Redundancy as an Opportunity. A key inefficiency in VR streaming is the duplication of chrominance information across stereo views. While luminance encodes the structural differences needed for depth perception, chroma is nearly identical between left and right views. Transmitting both views in full color wastes bandwidth without improving perceptual quality. Exploiting this redundancy opens a new design direction: transmit one view in color and the other in grayscale, reconstructing the missing chroma on the client-side. This observation underpins the design of iSR, which integrates colorization and super-resolution to reduce bandwidth requirements while preserving high visual fidelity.

General-Purpose Super-Resolution Models. Single-image super-resolution (SISR) methods such as EDSR [25] and the transformer-based SwinIR [23] achieve strong perceptual quality on natural images. However, these models are designed for offline 2D image

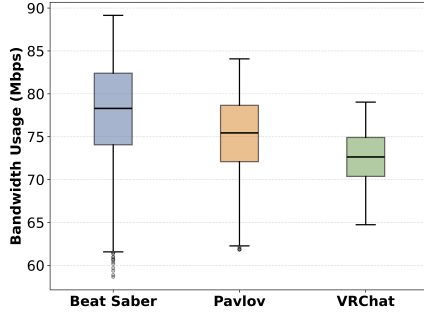


Figure 2: Bandwidth usage distribution for three VR games.

enhancement and typically require substantial computational overhead, which makes them unsuitable for real-time VR client devices. We discuss these constraints further in §4.9.

Super-Resolution for Gaming. In the gaming industry, SR has been widely adopted for gaming through solutions such as NVIDIA’s DLSS [28], AMD’s FidelityFX Super Resolution (FSR) [3], and Intel’s Xe Super Sampling (XeSS) [18]. These methods combine advanced upsampling with perceptual priors, allowing gaming engines to render at lower resolutions and then upscale frames to near-native quality. This saves processing power and allows running complex games on lower-end GPUs. Recently, VR-specific extensions for these SR methods have been introduced, such as DLSS 2.1, which accounts for stereo rendering and head-tracked motion. In addition, several works in the literature have proposed SR methods for 2D and VR games, including [21, 41, 47].

Most of the aforementioned SR methods, however, are designed for *stand-alone* gaming, not cloud gaming. This is because they require tight integration with the game source code and accessing data structures internal to the game renderer, such as depth buffers and motion vectors [33]. In cloud gaming, the renderer runs on the cloud, not at the client side. Thus, such data structures are not available at the client side for the SR methods to access. In addition, running the SR methods on the cloud would defeat the purpose of using it as the full-resolution encoded frames would have to be streamed to the client.

Super-Resolution for VR Applications. Beyond 2D gaming, several works have specifically targeted SR for immersive VR and 360° content. For example, [19] proposes S3PO, a deep learning model tailored for omnidirectional video SR that introduces a 360° video dataset (360VDS) and addresses spherical distortions inherent in equirectangular projections. Similarly, [47] designs CVHSSR, a cross-view hierarchy network for stereo SR that leverages intra-view hierarchies and cross-view attention, demonstrating SOTA quality on stereo benchmarks but with high complexity that limits practical VR streaming. These efforts highlight the promise of VR-specific SR (details provided in §4.9), but also underscore the challenge of balancing perceptual fidelity with the strict latency and efficiency requirements of real-time VR applications.

Foveated Rendering and Encoding. Foveated rendering exploits the non-uniform acuity of the human visual system by rendering the foveal region at high resolution while reducing detail in the periphery. Variants such as static, gaze-tracked [9], multi-resolution,

content-aware [11], and asymmetric [36] methods have demonstrated substantial reductions in rendering cost. Foveated encoding applies the same principle at the video compression stage, allocating more bits to the fovea and fewer to peripheral regions [2, 15, 17]. Foveated approaches, however, rely on accurate eye tracking, require specialized hardware, and are sensitive to sudden gaze shifts, which limit their wide deployability in practice.

3 Proposed Solution

In this section, we first summarize the design goals of iSR. Then, we present an overview of its operation and main components, followed by the details of each component.

3.1 Design Goals

Our goal is to design a bandwidth-efficient system for cloud VR gaming that minimizes transmission cost while maintaining an immersive experience. To achieve this, we design iSR to achieve the following goals:

- (1) **Substantial Bandwidth Reduction:** The primary objective is to significantly reduce the required bandwidth to support high-resolution VR streaming even under constrained bandwidth, without compromising the user experience. By lowering transmission cost, iSR enables smooth gameplay and consistent visual quality across diverse network conditions.
- (2) **Visual Quality Preservation:** Maintaining visual fidelity is essential to sustain immersion in VR environments. iSR is designed to minimize perceptual degradation and deliver high-quality visuals faithful to the original content.
- (3) **Low Latency and Computational Efficiency:** VR cloud gaming is highly latency-sensitive, and excessive overhead can degrade the user experience. iSR limits additional computational cost on both server and client, making it feasible to integrate iSR into existing cloud VR pipelines without introducing noticeable delays.
- (4) **Agnostic to Game Engine Data Structures:** VR cloud gaming renders games on the servers. Thus, all rendering-related data structures are not available on the client side. To enable deployment in VR cloud gaming systems where the clients perform upscaling, iSR is designed to operate independently of game engines, internal game data structures, and metadata. This also simplifies deployment and avoids compatibility issues with proprietary game engines and assets.
- (5) **No Modifications in Game Source Code:** The source code of most games is proprietary. Thus, upscalers that require changing game source code suffer from limited deployability, since they impose additional costs. In addition, for legacy but still popular games, the code may not be available or easily modifiable. iSR does not need to access or change the game source code or perform any custom integration for individual games; it operates on the encoded video streams of the games. Thus, it supports a wide range of recent and legacy VR games without requiring additional software development efforts, thereby minimizing cost and time while offering wide deployability.

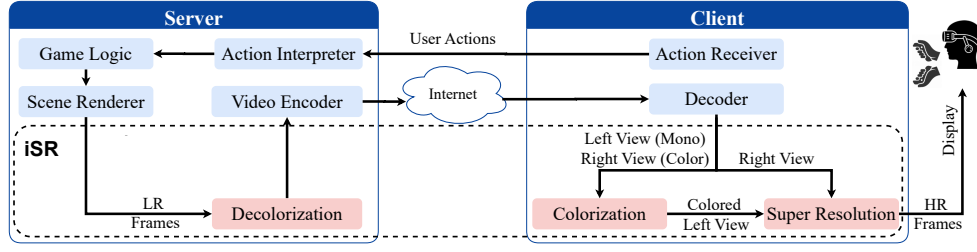


Figure 3: Overview of the proposed iSR system for reducing the required bandwidth in VR cloud gaming platforms.

3.2 System Overview

Figure 3 provides an overview of iSR and its main components. On the server side, player actions are interpreted and processed by the game logic to generate corresponding in-game events. The scene renderer produces stereo frames representing these events. The frames are rendered at low resolution (LR) to conserve computing and bandwidth resources. The left view is then decolorized by removing its chrominance and retaining only the luminance, thereby eliminating redundant color information across views and further reducing the required bandwidth. The LR frames are finally compressed by the video encoder before being streamed over the network to the client.

On the client side, the received frames are decompressed by the video decoder. A colorization module reconstructs the missing color in the left view using the color information from the right view. Then, a super-resolution module upsamples both views to high-resolution (HR) quality. The reconstructed views are then merged to recreate the final VR stereo frames for display.

A notable advantage of iSR is its seamless integration with existing commercial cloud VR pipelines, such as CloudXR, Shadow PC, and Virtual Desktop, without requiring any modifications to their rendering or compositor stages. Server-side streaming remains unchanged apart from a lightweight decolorization step applied prior to standard encoding such as H.264, H.265, and AV1. Client-side reconstruction runs as a GPU post-processing stage after decoding. This architecture therefore supports practical integration into existing cloud VR pipelines.

The operation of iSR can be summarized as follows:

$$(L_{LR}^{Mono}, R_{LR}^{Color}) \xrightarrow{iSR} (\hat{L}_{HR}^{Color}, \hat{R}_{HR}^{Color}) \quad (1)$$

In the above equation, the left view (L) is transmitted in monochrome and reconstructed to full color on the client, while the right view (R) retains color throughout the pipeline. The symbols $\hat{L}$ and $\hat{R}$ denote the reconstructed outputs of iSR. This design significantly reduces the required bitrate while preserving perceptual quality, offering a practical and efficient solution for VR cloud gaming. We provide details of the colorization and super-resolution components in the following sections.

3.3 Details of iSR

The goal of our work is to minimize the required bandwidth of VR cloud gaming, *without* significantly impacting the visual quality of the displayed frames *and* within the strict latency requirements of interactive games. To achieve this goal, we first propose a lightweight colorization module for VR cloud gaming, which removes

the redundant color information in the stereo frames at the server side. Then, on the client side, we restore the missing color and integrate an efficient upscaling module to enhance spatial resolution.

Potential Gain from Removing Colors. Modern video coding standards encode video in the YUV color space and apply chroma subsampling within individual video frames. In this representation, each pixel can have one luminance (Y) and two chrominance (U, V) components. Chroma subsampling means that the chrominance components may not exist in all pixels. This is done to save memory and processing resources.

To illustrate chroma subsampling, consider 4 pixels arranged in a 2×2 block. There are various ways for chroma subsampling, including the following examples:

- YUV444: No subsampling; All 4 pixels have Y, U, and V.
- YUV422: Chroma components are halved horizontally; two pixels share U and V.
- YUV420: Chroma components are halved horizontally and vertically; the 2×2 block shares the same U and V.
- YUV400: Luminance only, no chroma components.

There are two frames in stereo VR streams. If we use the YUV444 format for both frames, for example, each pixel would have 3 YUV samples in the left view and another 3 samples in the right view. That is, the cost per pixel is 6 (multiplied by the cost of each of the YUV components). In contrast, if we were to use the YUV444 format for the left view and the YUV400 (no chroma components) format for the right view, the cost per pixel would drop to 4, resulting in potential savings of 33.3%.

key Insight. We observe that luminance in stereo views encodes the structural differences necessary for depth perception, whereas chrominance is nearly identical between the two views. That is, there is significant redundancy in the chrominance components in stereo views. Based on this observation, we propose sending the chrominance components of only one of the stereo views and reconstructing the chrominance of the other view at the client side.

Potential Savings. We calculate the potential bandwidth savings that can be achieved by transmitting the chrominance components of only one view for various chroma subsampling formats and present the results in Table 1. The table indicates potential savings ranging from 16.7% to 33.3%, which are significant.

It is important to note that these values represent raw sample counts before compression. In practice, codecs such as H.265 further benefit from the lower entropy of monochrome frames, often yielding savings beyond the theoretical percentages. We further note that these bandwidth savings come at the cost of performing

Table 1: Potential bandwidth savings due to sending the chrominance components of only one view for various chroma subsampling formats.

Right View + Left View	Drop in Pixel Cost	Bandwidth Saving
YUV444 + YUV400	6.0 → 4.0	33.3%
YUV422 + YUV400	4.0 → 3.0	25.0%
YUV420 + YUV400	3.0 → 2.5	16.7%
YUV444 + YUV444	6.0 → 6.0	0.0%

colorization on the client side, which must be efficient and done in real-time.

In summary, removing redundant chroma from one of the stereo views reduces the *raw* per-pixel cost. For the widely used YUV420 format, the savings are 16.7%, while other formats, such as YUV422 and YUV444, can achieve even higher savings. Moreover, practical codecs (e.g., H.265/HEVC) benefit from the lower entropy of monochrome frames and can achieve greater savings. Achieving such savings requires accurate reconstruction of the removed chroma, which motivates the need to identify the most effective colorization approach.

Colorization in Prior Work. Traditional colorization techniques rely on heuristics or user input, such as scribbles [42], which are infeasible for real-time VR streaming. More recent CNN-based methods can automatically hallucinate plausible colors [22, 35], but they do not leverage reference images and therefore lack the accuracy and stereo consistency required for VR. Reference-based approaches transfer chroma from an external image to a grayscale target [46], which aligns better with VR since one stereo view can serve as a reference for the other. However, prior works generally assume unrelated scenes between the reference and target, leading to inaccuracies under stereo settings. This gap motivates our stereo-aware design tailored for VR game streams.

Proposed Lightweight Colorization Module. The architecture of the proposed colorization module is illustrated in Figure 4, which extends the SOTA *general* colorization method, SiCNet [45], and makes it run in real time for VR cloud gaming.

As the most recent stereo colorization framework, SiCNet surpasses prior approaches in terms of color fidelity and structural consistency on natural image datasets. VR stereo imagery presents a comparable scenario: the two views share nearly identical chroma while differing primarily in structure. This analogy motivated us to adopt stereo-aware colorization for VR bandwidth optimization, where one view is transmitted in monochrome and reconstructed using the chroma information from the other view.

While SiCNet achieves high accuracy on natural image datasets and generalizes well to real-world stereo data, our analysis reveals several limitations when applied to VR gaming, including poor perceptual fidelity on dynamic game textures, high inference latency, and high memory usage. We address these issues with three optimizations.

The first optimization is *module pruning*. We remove the Stereo Fusion Module (SFM) and Stereo Consistency Module (SCM) as they provide only marginal quality improvements in VR but introduce inference overhead. The second optimization is the use of *efficient convolutions*. Standard 3×3 convolutions are replaced with depthwise-separable convolutions, which significantly reduce

the computational cost of convolutional layers while maintaining perceptual quality with minimal loss.

The third optimization is *quantization*. We apply post-training quantization to 8-bit precision for both weights and activations. This reduces both memory usage and latency, while preserving perceptual quality across stereo frames. The impact of these three optimizations, along with the effect of feature-channel width on the trade-off between quality and efficiency, is analyzed through ablation studies in §4.

Enhancing Resolution of VR Gaming Streams. To restore high-resolution quality of the VR streams, iSR employs a recent upscaling model, GameSR [31], which is designed for 2D games and computationally efficient. While many SR models exist, most are either too computationally heavy (e.g., EDSR [25], IMDN [16], LatticeNet [26], and SwinIR [23]) or optimized for natural images rather than gaming content. GameSR strikes a balance between efficiency and perceptual quality, preserving edges and textures that are critical for immersion in VR while keeping runtime and memory requirements modest. Its residual-convolutional architecture, combined with pixel-shuffle upsampling, makes it more efficient than transposed convolution-based alternatives that often introduce artifacts.

Beyond spatial fidelity, GameSR integrates a lightweight temporal component that enhances reconstruction across frames. This helps ensure smooth detail propagation during fast gameplay, a property particularly important in VR systems where magnified pixels make temporal artifacts highly distracting.

In our pipeline, the right view is transmitted in full color while the left is sent in grayscale. After colorization restores the missing chroma, both views are passed to the upscaling module, which reconstructs them to the original resolution. This sequential design offers two benefits: (1) Reduces transmission bitrate by operating on low-resolution inputs, and (2) Lowers the computational cost of colorization as fewer pixels need to be processed before upscaling.

4 Evaluation

In this section, we evaluate iSR in terms of bandwidth reduction, perceptual quality, and computational efficiency. The experiments reflect a practical cloud VR pipeline where the right view is transmitted in full color and enhanced by super-resolution. In contrast, the left view is transmitted in monochrome and reconstructed through both colorization and super-resolution. This setup highlights the core challenge of maintaining visual fidelity under reduced chroma transmission. The results demonstrate how iSR achieves substantial bitrate savings while remaining within the latency and resource budgets of VR cloud gaming.

4.1 Datasets and Experimental Setup

Datasets.

To evaluate our method, we collected and curated a dataset consisting of VR games with diverse genres, visual styles, and levels of complexity to reflect representative gameplay conditions. This selection ensures that iSR is tested under a broad range of scenarios, including fast-paced motion, high-detail textures, stylized environments, and unpredictable visual dynamics. The games selected for this evaluation are as follows:

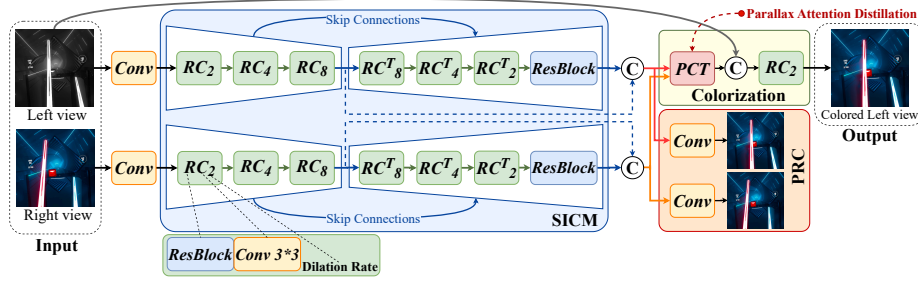


Figure 4: Architecture of the proposed lightweight colorization module designed for interactive VR cloud gaming systems.

- **Beat Saber:** A fast-paced rhythm game where players slice moving colored blocks to music. Its rapid motion and color-critical gameplay make it ideal for testing temporal coherence, color accuracy, and responsiveness under bandwidth constraints.
- **Pavlov VR:** A competitive first-person shooter game often compared to Counter-Strike. It features diverse indoor and outdoor maps with frequent viewpoint shifts, weapon effects, and varied lighting, providing a challenging benchmark for robustness under visually complex and dynamic conditions.
- **VRChat:** A social VR platform with highly diverse, user-generated worlds. Its environments range from photorealistic rooms to stylized cartoon-like scenes, making it valuable for evaluating generalization across different visual styles and texture statistics.

Since no public dataset contained stitched stereo VR frames for these games, we collected our own using a Meta Quest 3 headset at its native resolution of $4,128 \times 2,208$ pixels. Recordings were captured through the built-in screen recorder after configuring capture settings via Android Debug Bridge (adb) [4], a command-line interface for controlling Android-based devices. Videos were transferred and converted into individual frames to construct the dataset.

Experimental Setup. All experiments were conducted on a Linux workstation (Ubuntu 22.04) with an Intel® Xeon® w7-2495X (32 GB RAM), and an NVIDIA RTX A4000 GPU (20 GB VRAM). The implementation was developed using PyTorch 2.5.1 and accelerated with CUDA 11.8 and cuDNN 9.1. Using the dataset we collected and the described experimental setup, we evaluate the performance of our proposed system across different categories of metrics discussed in the following sections.

4.2 Evaluation of iSR

In this section, we evaluate the end-to-end performance of iSR in terms of bandwidth reduction, perceptual quality, and computational efficiency. In each experiment, the colorization module reconstructs the chroma of the left view using the right view as a reference, after which the super-resolution module restores both views to high-resolution stereo frames. The right view is enhanced by super-resolution only, whereas the left view requires both colorization and super-resolution, making it the more challenging case. This setup reflects the practical cloud VR pipeline, where only one view is transmitted in full color. It demonstrates how the two modules jointly enable bandwidth savings while maintaining visual fidelity and low resource overhead.

4.3 Rate-Distortion (RD) Analysis for VR

We first analyze the rate-distortion (RD) behavior of our VR pipeline at the native 4K resolution. Raw VR streams were downsampled and encoded at lower resolutions, then decoded to mimic client reception under constrained bandwidth. These low-resolution versions were subsequently reconstructed to the native resolution using our upscaling module with $4\times$ and $8\times$ scaling, while the unmodified 4K streams serve as the baseline. Following prior work, we categorize perceptual quality using VMAF thresholds: *fair/watchable* at ≥ 70 [5], *good* at ≥ 80 , and *excellent* at ≥ 90 [10, 30].

As shown in Figure 5, $4\times$ scaling provides the most balanced trade-off. It reaches *watchable* quality at only 8 Mbps, compared to 18 Mbps for the baseline, resulting in a bitrate savings of 56%. For *good* quality, $4\times$ scaling requires just 12 Mbps, whereas the baseline demands 24 Mbps, corresponding to a 50% reduction. At higher bitrates, $4\times$ scaling continues to improve, reaching VMAF ≈ 90 at 26 Mbps, which we consider in the *excellent* category, while the baseline never reaches this threshold in the 2-26 Mbps range. By contrast, $8\times$ scaling, as shown in Figure 6, delivers *watchable* quality at 12 Mbps, while the baseline remains far below. However, the curve saturates quickly, plateauing around VMAF 73 as the network provides insufficient detail for further reconstruction. Similar trends are observed in PSNR and SSIM, where $4\times$ scaling consistently outperforms the baseline across the bitrate range, while $8\times$ scaling improves low-bitrate performance but saturates early.

In summary, $8\times$ scaling is most suitable as an emergency mode under severe bandwidth constraints, whereas $4\times$ scaling emerges as the practical configuration for VR streaming, substantially shifting the RD curve leftward and enabling high-quality experiences at nearly half the bitrate of native 4K transmission.

4.4 End-to-end Latency

To quantify latency in a realistic cloud VR setup, we used GameLab [1] a python-based real-time cloud gaming testbed. Due to a combination of system and network constraints, and the lack of stable 4K support in our current aiortc-based pipeline, results in this section stream 2K (2560×1440) stereo content; our 4K results are therefore restricted to offline rate-distortion in Section 4.3.

Figure 7 reports the end-to-end latency breakdown for our VR streaming setup using the most aggressive configuration, *iSR* $\times 8$, where the client reconstructs $8\times$ upscaled stereo views from a downsampled color view and a downsampled monochrome view. Across

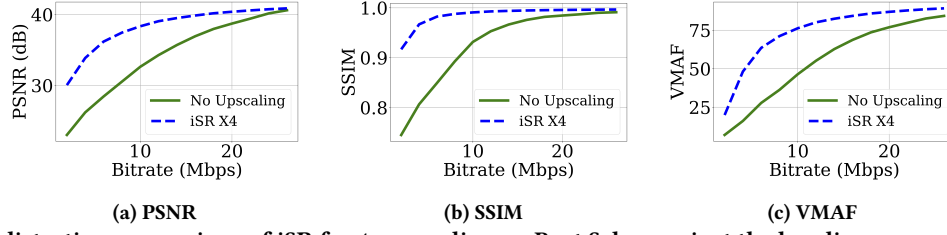


Figure 5: Rate-distortion comparison of iSR for 4× upscaling on Beat Saber against the baseline across different bitrates.

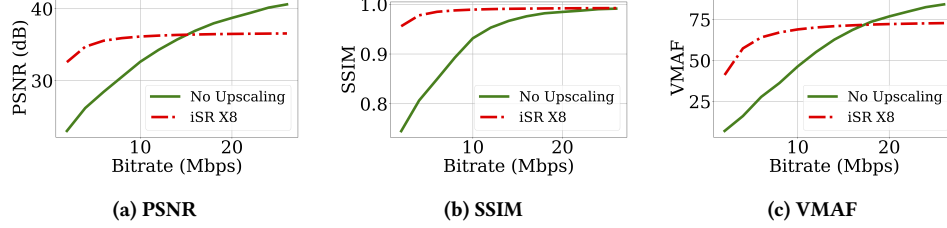


Figure 6: Rate-distortion comparison of iSR for 8× upscaling on Beat Saber against the baseline across different bitrates.

Table 2: Perceptual quality of iSR across three VR games and all scales.

Scale	Beat Saber			Pavlov VR			VRChat		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
×4	39.50	0.9835	0.0128	40.23	0.9789	0.0154	37.18	0.9701	0.0283
×6	34.52	0.9659	0.0581	36.95	0.9598	0.0642	33.40	0.9565	0.0580
×8	34.49	0.9674	0.0472	35.72	0.9531	0.0718	32.70	0.9490	0.0705
×10	34.46	0.9562	0.0559	34.98	0.9447	0.0823	32.30	0.9408	0.0820

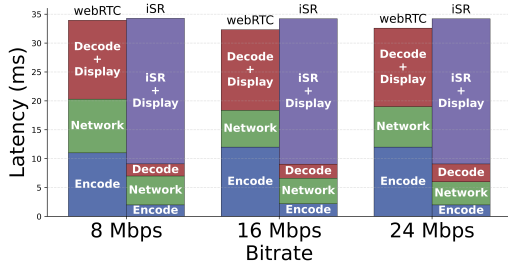


Figure 7: End-to-end latency breakdown across bitrates for WebRTC and iSR.

8, 16, and 24 Mbps, baseline WebRTC attains total encode-network-decode-display latencies of 33.95, 32.33, and 32.59 ms, while iSR remains tightly clustered around 34 ms (34.29, 34.22, and 34.22 ms), i.e., within roughly 1-2 ms of the baseline despite performing stereo-aware colorization and 8× super-resolution on the client. The main difference lies in how this budget is allocated: because iSR streams a downsampled color view plus a monochrome view, its *encode + network* component shrinks by roughly 2.5–3× compared to WebRTC, while the client-side processing segment grows to account for colorization, upscaling, and display. Nevertheless, the overall end-to-end latency stays well below typical VR streaming limits.

4.5 Perceptual Quality

We evaluate the perceptual quality of iSR on the curated datasets introduced in §4.1, focusing primarily on Beat Saber while also including Pavlov VR and VRChat to confirm generalizability across different genres. These three games offer complementary challenges: Beat Saber stresses rapid motion and color-critical interactions; Pavlov VR features diverse textures, complex lighting, and frequent viewpoint changes; and VRChat spans a wide range of user-generated environments, from photorealistic rooms to highly stylized cartoon-like worlds. Together, they form a robust testbed for assessing reconstruction fidelity under varied VR conditions.

We report standard distortion metrics: PSNR, which quantifies pixel-wise fidelity, and SSIM, which evaluates structural similarity by modeling luminance, contrast, and texture preservation. To complement these, we use LPIPS [43], a learned perceptual metric that correlates more closely with human visual judgments by comparing deep feature representations rather than raw pixels. Together, these metrics provide a balanced assessment of both signal-level accuracy and perceived visual quality.

As shown in Table 2, iSR maintains strong performance across scale factors and genres. For Beat Saber, the system achieves a PSNR of up to 39.5 and an SSIM above 0.98 at a ×4 scaling factor, with graceful degradation at higher factors, where bitrate savings are most pronounced. LPIPS values remain low (≈ 0.05 or lower), indicating high perceptual similarity to ground truth. Pavlov VR follows similar overall trends but with slightly lower SSIM and higher

Table 3: MOS for iSR across all VR games and scale factors.

Game	×4	×6	×8	×10
Beat Saber	4.88	4.50	4.00	3.75
Pavlov VR	4.75	4.38	3.88	3.63
VRChat	4.25	3.88	3.63	3.13

LPIPS, which is attributed to its more complex and dynamic scenes; however, PSNR values remain in the good quality range. VRChat exhibits the largest variability due to its diverse user-generated environments, yet iSR still achieves strong perceptual fidelity, with SSIM values above 0.94 and LPIPS below 0.03–0.08 across scales. These results demonstrate that iSR generalizes well from stylized rhythm gameplay to realistic first-person scenes and even to highly heterogeneous VRChat environments.

Beyond quantitative results, we also examine reconstruction fidelity qualitatively. Figure 8 presents representative stereo frames from Beat Saber, Pavlov VR, and VRChat at the ×8 scale, where the left eye undergoes both colorization and super-resolution. To better highlight fine-grained differences, the figure includes ×4 magnified insets of the left view. The reconstructed outputs closely resemble the ground truth and preserve spatial detail and color fidelity across both fast-paced and visually complex content. Even under aggressive downsampling, iSR avoids distracting artifacts, demonstrating robustness under bandwidth-limited scenarios.

User Study. To further evaluate perceptual quality, we conducted a user study using recorded gameplay sessions upscaled by iSR. A total of 8 participants took part, each viewing representative clips from the three VR games (Beat Saber, Pavlov VR, and VRChat) across all four scaling factors. Among the participants, 37.5% were experienced gamers, 37.5% were occasional gamers, and 25% had little or no prior gaming experience.

After each viewing, participants rated the visual quality on a 5-point Mean Opinion Score (MOS) scale. The results, summarized in Table 3, demonstrate consistently high perceptual quality across all games and scale factors. iSR achieved an average MOS of 4.88/5 for Beat Saber, 4.75/5 for Pavlov VR, and 4.25/5 for VRChat at the ×4 scale. Even at higher scaling factors such as ×8 and ×10, iSR maintained perceptual quality above 3.6 for all games. These findings validate iSR’s ability to preserve high-quality user experience under aggressive bandwidth reduction, complementing our objective metrics.

Taken together, these findings demonstrate that iSR achieves perceptually faithful reconstruction across different VR genres and scaling factors. By combining stereo-aware colorization with efficient SR, it delivers high-quality VR experiences that remain well within the latency and bandwidth constraints of cloud gaming.

4.6 Inference Time and Memory Requirements

Table 4 summarizes the runtime and memory usage of iSR across different upscale factors. The results reveal a consistent trend: as the upscale factor increases, inference becomes both faster and lighter in memory. This occurs because larger upscale factors operate on proportionally smaller low-resolution inputs, reducing the computational and memory requirements of the network.

At ×6, iSR runs at 32.9 ms per frame, which already falls within the 50 ms threshold generally considered acceptable for VR. The

Table 4: Inference time and memory usage of iSR.

Scale	Time per frame (ms)	Mem (GB)
×4	73.61	2.310
×6	32.92	0.932
×8	24.38	0.624
×10	17.68	0.538

Table 5: Impact of SiCNet modifications across all scales.

Scale	Variant	PSNR↑	SSIM↑	Inf (ms)↓	Mem (GB)↓
×4	Baseline	41.10	0.9869	123.0	3.0
	- SFM/SCM	40.50	0.9863	114.2	2.7
	+ Depthwise Conv	40.28	0.9846	109.5	2.3
	+ Quantization	40.12	0.9823	66.8	2.0
×6	Baseline	39.70	0.9852	41.0	1.298
	- SFM/SCM	39.49	0.9848	35.9	1.128
	+ Depthwise Conv	38.22	0.9822	34.3	0.932
	+ Quantization	38.71	0.9821	22.9	0.778
×8	Baseline	40.56	0.9852	35.3	0.724
	- SFM/SCM	40.16	0.9848	32.2	0.648
	+ Depthwise Conv	39.99	0.9822	26.3	0.550
	+ Quantization	39.81	0.9821	19.7	0.488
×10	Baseline	34.92	0.9713	6.5	0.566
	- SFM/SCM	34.05	0.9710	6.0	0.508
	+ Depthwise Conv	33.89	0.9705	5.7	0.458
	+ Quantization	33.71	0.9705	5.3	0.396

×8 and ×10 configurations further reduce runtime to 24.4 ms and 17.7 ms, respectively, with the latter even meeting the stricter 20 ms requirement to prevent motion sickness. These findings demonstrate that iSR can meet the latency demands of VR across multiple upscale factors while maintaining efficient memory usage.

4.7 Analysis of the Colorization Module

We conduct a detailed analysis of the proposed lightweight colorization module, focusing on both reconstruction quality and computational efficiency. Since colorization is applied to every frame in real time, its runtime and memory consumption are as important as its perceptual accuracy. So, we evaluate the effect of architectural modifications, channel width, and perceptual fidelity using standard quality metrics as well as perceptual color difference measures.

Effect of Architectural Modifications. We begin by analyzing the impact of successive modifications made to adapt the baseline stereo colorization model (SiCNet) for VR cloud gaming. Table 5 reports results across all scaling factors. Removing the Stereo Fusion and Stereo Consistency modules slightly reduces quality but lowers inference time and memory usage. Replacing standard convolutions with depthwise-separable convolutions further reduces latency (by nearly 20%) while maintaining PSNR and SSIM. Finally, applying post-training quantization halves the runtime with negligible perceptual loss. Together, these modifications significantly reduce inference time and memory footprint across scales, making the module suitable for interactive VR.

Impact of Feature Channels. One of the challenges in adapting deep models for VR streaming is striking a balance between reconstruction quality and computational efficiency. Higher channel

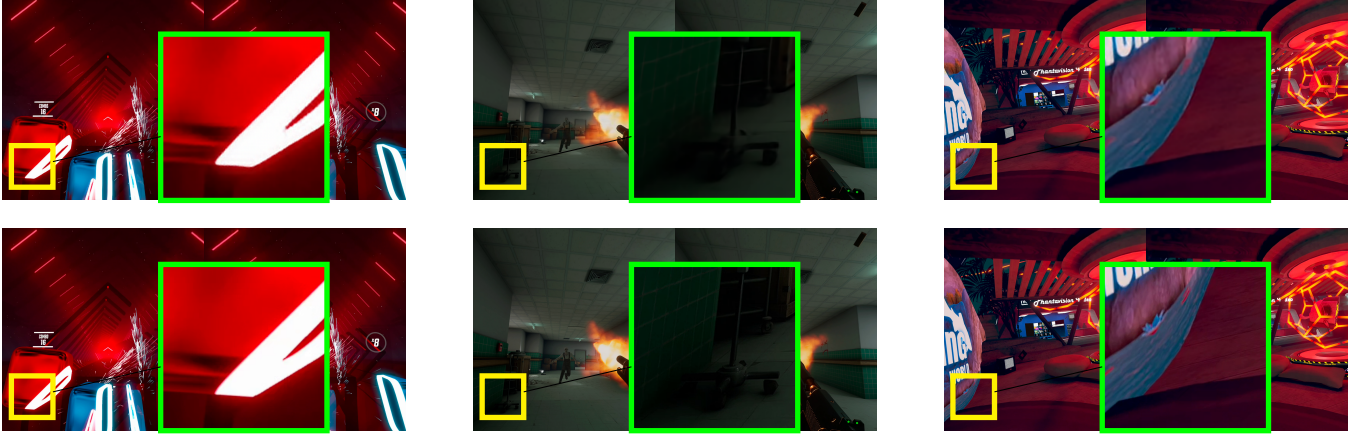


Figure 8: Demonstrating the visual quality produced by iSR with $8\times$ bandwidth reduction on sample left-eye frames from Beat Saber, Pavlov VR, and VRChat. The upscaled frames (top) are nearly indistinguishable from the ground truth (bottom).

widths in convolutional layers generally capture richer representations, which can improve fidelity, but they also increase inference time and memory usage. For VR cloud gaming, this trade-off is particularly critical. The system must deliver high perceptual quality under strict bandwidth and resource constraints.

To study this balance, we vary the number of feature channels in our colorization network from 16 to 64 and report results across all scale factors in Table 6. We did not include 64 channels at $\times 4$, as the inference time at 48 channels already reached 106 ms per frame, making higher widths clearly infeasible for VR. Increasing channel width improves quality, but at the cost of heavier computation. For example, at $\times 8$, moving from 16 to 48 channels raises PSNR by 7.66 and SSIM by 0.0186, but inference time more than doubles (9.44 ms $\rightarrow$ 19.70 ms) and memory usage grows by 32.8%. Compared to 32 channels, the 48-channel setting achieves only modest quality gains (+1.19 PSNR, +0.0021 SSIM) at the cost of a 52.4% increase in latency.

Balancing these factors, we select different widths depending on the scale: for $\times 4$ and $\times 6$, we adopt 32 channels as they provide strong reconstruction quality at substantially lower latency; for $\times 8$ and $\times 10$, we use 48 channels, which give a better trade-off between accuracy and efficiency at higher compression ratios. This configuration provides the most practical end-to-end balance across all operating points of iSR.

Perceptual Color Difference Evaluation. To directly assess perceptual color fidelity, we use ΔE_{2000} , a widely adopted metric in imaging and color science that measures perceptual differences between two colors in the CIELAB space [32]. Unlike signal-based metrics such as PSNR or SSIM, which quantify pixel-wise similarity or structural consistency, ΔE_{2000} is explicitly designed to model the nonlinear characteristics of human color perception. Small variations in some hues can be highly noticeable to viewers even if PSNR and SSIM remain high, and ΔE_{2000} accounts for these perceptual sensitivities. Lower values indicate closer color matches, with $\Delta E < 2.3$ considered imperceptible and $\Delta E < 5.1$ is generally regarded as acceptable in visual applications.

Figure 9 reports the cumulative distribution of ΔE across all scale factors. Our results show that over 85% of reconstructed images

Table 6: Impact of channel across all scales.

Scale	#Ch	PSNR $\uparrow$	SSIM $\uparrow$	Inf (ms) $\downarrow$	Mem (GB) $\downarrow$
$\times 4$	16	35.90	0.9751	38.91	2.016
	32	40.12	0.9823	66.61	2.340
	48	41.43	0.9844	106.28	2.720
$\times 6$	16	34.12	0.9722	15.70	0.778
	32	38.71	0.9821	22.88	0.932
	48	40.28	0.9846	34.68	1.128
	64	41.28	0.9857	48.39	1.298
$\times 8$	16	32.15	0.9646	9.44	0.488
	32	38.62	0.9811	12.93	0.550
	48	39.81	0.9832	19.70	0.648
	64	35.34	0.9797	25.40	0.724
$\times 10$	16	34.21	0.9712	6.85	0.396
	32	37.00	0.9804	8.66	0.458
	48	39.18	0.9830	10.53	0.508
	64	39.65	0.9837	13.51	0.566

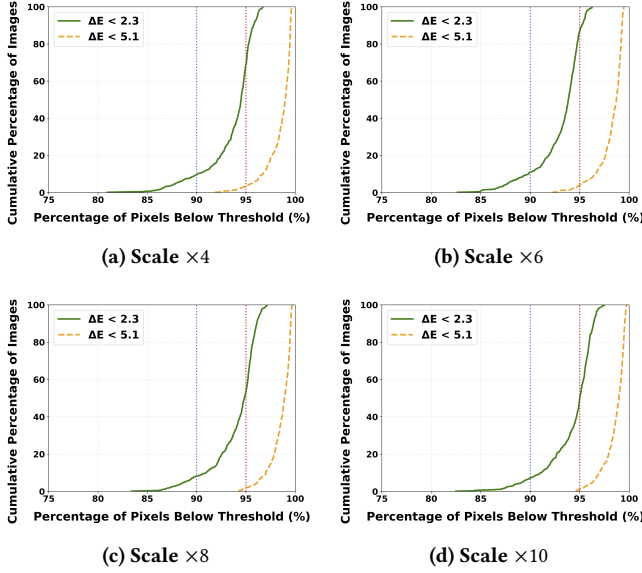
achieve $\Delta E < 2.3$, while more than 98% remain below $\Delta E < 5.1$. This confirms that our stereo-aware colorization delivers perceptually faithful results even under aggressive downsampling.

While stereo-aware colorization eliminates redundant chroma across views, the frames remain spatially downsampled to reduce bitrate. To restore immersive quality, the next stage applies super-resolution, which reconstructs high-resolution detail from low-resolution, colorful frames. Together, the two modules form a complementary pipeline: colorization reduces inter-view redundancy, and super-resolution counteracts quality loss from downsampling.

4.8 Comparison against SOTA Colorization

In this section, we present the results from our evaluation against the SOTA colorization model **SiCNet** [45].

We evaluated a pre-trained SiCNet model directly on our gaming content. The results for an upscaling factor of $\times 4$ are summarized in Table 7. While SiCNet delivers strong performance on natural scenes, we observed a substantial degradation when applied to gaming assets, with PSNR dropping to 21.02 and SSIM to 0.526. In contrast, our lightweight colorization achieves a PSNR of 39.78 and

Figure 9: Cumulative distribution of ΔE_{2000} across all scales.

SSIM of 0.983, demonstrating much better preservation of game textures and edges. This performance gap arises because SiCNet was trained primarily on natural stereo photographs, where color and depth relationships follow real-world statistics. In contrast, gaming content contains synthetic lighting, sharp edges, and non-photorealistic textures that differ significantly from natural images, causing poor generalization to this domain.

Regarding inference cost, we were unable to execute SiCNet on our largest GPU with 24 GB VRAM (NVIDIA RTX 3090) due to excessive memory demand, and CPU inference required approximately 8 s per frame, highlighting impractical deployment.

4.9 Comparison against SOTA SR

We compare iSR against CVHSSR [47], which is the SOTA SR model explicitly designed for VR content, as well as two widely used lightweight SR models, EDSR [24] and SwinIR [23]. For this experiment, we upsampled both the left and right views by a factor of 4. The results are summarized in Table 8. Despite CVHSSR’s excellent reconstruction quality (42.79 PSNR, 0.989 SSIM), its runtime of ~ 2947 ms per frame makes it unsuitable for *interactive* VR in cloud gaming. EDSR and SwinIR offer lower computational cost than CVHSSR, but their inference times (105 ms and 1898 ms per frame) remain far above the real-time requirements of cloud VR, and both models deliver noticeably lower perceptual quality (39.74 and 39.70 PSNR). In contrast, our lightweight approach achieves better perceptual quality (44.06 PSNR, 0.991 SSIM) while maintaining a real-time runtime of 12.87 ms per frame. These results emphasize the feasibility of our design compared to existing SOTA SR models for latency-sensitive VR streaming scenarios.

4.10 Colorization and Upscaling Trade-off

This section analyzes the relationship between visual quality, bitrate savings, and computational efficiency in iSR. Our goal is to identify scaling factors that reduce transmission cost while maintaining

Table 7: Colorization results ($\times 4$ upscaling).

Model	Inf (ms) ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	66.89	39.78	0.983	0.007
SiCNet	-	21.03	0.526	0.120

Table 8: Super-resolution results ($\times 4$ upscaling).

Model	Inf (ms) ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	12.87	44.06	0.991	0.005
CVHSSR	2946	42.79	0.989	0.007
EDSR	105	39.74	0.978	0.012
SwinIR	1898	39.70	0.979	0.013

Table 9: Super-resolution results across all scales.

Scale	PSNR ↑	SSIM ↑	Inf (ms) ↓
$\times 4$	44.06	0.9911	12.87
$\times 6$	36.29	0.9708	19.49
$\times 8$	36.01	0.9604	13.96
$\times 10$	36.71	0.9733	9.94

perceptual fidelity for cloud VR gaming. Table 9 shows that iSR preserves visual quality across all scales. At $\times 4$, it achieves the highest fidelity (PSNR ≈ 44 , SSIM > 0.99), while higher scales trade quality for efficiency. Notably, $\times 8$ and $\times 10$ still deliver SSIM above 0.96 with runtimes below 14 ms and 10 ms, respectively, making them attractive for low-bandwidth or latency-sensitive scenarios.

The bitrate–quality trade-off is evident in Figures 5 and 6. Larger scaling factors substantially reduce required bitrate but also lead to rapid saturation in PSNR, SSIM, and VMAF. In contrast, $\times 4$ provides smoother quality gains at higher bitrates, making it preferable when bandwidth is limited but visual fidelity remains important.

In summary, SR enables iSR to aggressively downsample stereo frames while keeping perceptual degradation modest. Combined with colorization, which removes redundant chroma information, iSR achieves complementary spatial and chroma bandwidth reductions, making it practical for real-world cloud VR deployment.

5 Conclusion and Future Work

In this work, we presented iSR, a deep learning framework designed to reduce the bandwidth demands of immersive cloud VR gaming while preserving perceptual fidelity. iSR exploits the redundancy between stereo views by transmitting one downsampled color view and one downsampled monochrome view. On the client side, the system reconstructs full-resolution stereo pairs by combining stereo-aware colorization with a lightweight upscaling model. Our experiments on Beat Saber, Pavlov VR, and VRChat demonstrate that iSR achieves substantial bitrate savings while maintaining high perceptual quality. Importantly, the system operates efficiently within VR’s latency budget, confirming its deployability for bandwidth-constrained streaming environments.

Looking ahead, several promising directions remain. A possible next step is to unify colorization and upscaling into a single joint neural architecture, which could reduce computations, improve stereo consistency, and lower inference time.

References

- [1] 2026. GameLab: An AI-enabled cloud gaming testbed (source code). GitHub repository. <https://github.com/haseebfazzal/ai-testbed>
- [2] Ahmad Alhailal, Ze Wu, Yuk Hang Tsui, and Pan Hui. 2024. FovOptix: Human Vision-Compatible Video Encoding and Adaptive Streaming in VR Cloud Gaming. In *Proceedings of ACM Multimedia Systems Conference (MMSys'24)*. Bari, Italy, 67–77. doi:10.1145/3625468.3647612
- [3] AMD. 2022. AMD Fidelity Super Resolution. <https://www.amd.com/en/technologies/fidelityfx-super-resolution>
- [4] Android Developers. 2025. Android Debug Bridge (adb). <https://developer.android.com/tools/adb>
- [5] Netflix Technology Blog. 2018. VMAF: The Journey Continues. <https://netflixtechblog.com/vmaf-the-journey-continues-44b51ee9ed12>. Accessed: 2025-09-22.
- [6] Miguel Casanovas, Costas Michaelides, Marc Carrascosa-Zamacois, and Boris Bellalta. 2024. Experimental evaluation of interactive Edge/Cloud Virtual Reality gaming over Wi-Fi using unity render streaming. *Computer Networks* 243 (2024). doi:10.5281/zenodo.10455827
- [7] Kar Long Chan, Kohei Ichikawa, Yasuhiro Watashiba, and Hajimu Iida. 2017. Cloud-Based VR Gaming: Our Vision on Improving the Accessibility of VR Gaming. In *2017 International Symposium on Ubiquitous Virtual Reality (ISUVR)*. Nara, Japan, 24–25. doi:10.1109/ISUVR.2017.15
- [8] Malleshham Dasari, Arani Bhattacharya, Santiago Vargas, Pranjal Sahu, Aruna Balasubramanian, and Samir R Das. 2020. Streaming 360-Degree Videos Using Super-Resolution. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*. IEEE, Beijing, China, 1385–1394.
- [9] Farhaan Ebadulla, Chirag Mudlapur, and Gaurav BV. 2025. GazeProphet: Software-Only Gaze Prediction for VR Foveated Rendering. (8 2025). <http://arxiv.org/abs/2508.13546>
- [10] Elecard. 2023. Interpretation of Metrics: PSNR, SSIM, VMAF. https://www.elecard.com/page/article-interpretation_of_metrics. Accessed: 2025-04-23.
- [11] Linus Franke, Laura Fink, and Marc Stamminger. 2025. VR-Splatting: Foveated Radiance Field Rendering via 3D Gaussian Splatting and Neural Points. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 8, 1 (5 2025). doi:10.1145/3728302
- [12] Grand View Research. 2024. Virtual Reality in Gaming Market Size, Share & Trends Analysis Report By Component, By Region, And Segment Forecasts, 2024 - 2030. <https://www.grandviewresearch.com/industry-analysis/virtual-reality-in-gaming-market>
- [13] Grigore C. Burdea and Philippe Coiffet. 2003. *Virtual Reality Technology* (2 ed.). Wiley-IEEE Press.
- [14] Qiqi Hou, Farzad Farhadzadeh, Amir Said, Guillaume Sautiere, and Hoang Le. 2024. Low-Latency Neural Stereo Streaming. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, Washington, USA, 7974–7984.
- [15] Luke Hsiao, Brooke Krajancich, Philip Levis, Gordon Wetzstein, and Keith Winstein. 2018. Towards Retina-Quality VR Video Streaming: 15 ms Could Save Bandwidth. *ACM SIGCOMM Computer Communication Review* 48 (2018), 13–24. doi:10.1145/3274828.3274832
- [16] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiemei Wang. 2019. Lightweight Image Super-Resolution with Information Multi-Distillation Network. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2024–2032. doi:10.1145/3343031.3350905
- [17] Gazi Karam Illahi, Thomas Van Gemert, Matti Siekkinen, Enrico Masala, Antti Oulasvirta, and Antti Ylä-Jääski. 2020. Cloud Gaming with Foveated Video Encoding. *ACM Transactions on Multimedia Computing, Communications and Applications* 16, 1 (3 2020). doi:10.1145/3369110
- [18] Intel. 2024. Intel Xe Super Sampling API Developer Guide. <https://www.intel.com/content/www/us/en/developer/articles/guide/xe-super-sampling-developer-guide.html>
- [19] Qing Jiang et al. 2022. Omnidirectional Video Super-Resolution Using Deep Learning. *IEEE Transactions on Multimedia* 24 (2022), 4325–4338. doi:10.1109/TMM.2022.3151899
- [20] Jaehong Kim, Youngmok Jung, Hyunho Yeo, Juncheol Ye, and Dongsu Han. 2020. Neural-Enhanced Live Streaming: Improving Live Video Ingest via Online Learning. In *Proceedings of ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM'20)*, 107–125. doi:10.1145/3387514.3405856
- [21] Jia Li, Ziling Chen, Xiaolong Wu, Lu Wang, Beibei Wang, and Lei Zhang. 2024. Neural Super-Resolution for Real-time Rendering with Radiance Demodulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, Washington, USA, 4357–4367. <https://arxiv.org/abs/2308.06699>
- [22] Yifan Li, Yuhang Bai, Shuai Yang, and Jiaying Liu. 2024. COCO-LC: Colorfulness Controllable Language-based Colorization. In *MM 2024 - Proceedings of the 32nd ACM International Conference on Multimedia*. Association for Computing Machinery, Inc, New York, NY, USA, 10939–10947. doi:10.1145/3664647.3680662
- [23] Jingyun Liang, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. SwinIR: Image Restoration Using Swin Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 1833–1844. doi:10.1109/ICCVW54120.2021.00210
- [24] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced Deep Residual Networks for Single Image Super-Resolution. *CoRR* abs/1707.02921 (2017). <http://arxiv.org/abs/1707.02921>
- [25] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 136–144.
- [26] Zhengxiong Luo, Yan Huang, Liang Yu, Lizhuang Wang, Yulun Zhang, and Tatsuya Harada. 2020. LatticeNet: Towards Lightweight Image Super-Resolution with Lattice Block. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 272–289. doi:10.1007/978-3-030-58529-7_17
- [27] Costas Michaelides, Miguel Casanovas, Daniele Marchitelli, and Boris Bellalta. 2023. Is Wi-Fi 6 Ready for Virtual Reality Mayhem? A Case Study Using One AP and Three HMDs. In *Proceedings - 2023 IEEE Future Networks World Forum: Future Networks: Imagining the Network of the Future, FNWF 2023*. Institute of Electrical and Electronics Engineers Inc., Baltimore, Maryland, USA., 1–6. doi:10.1109/FNWF58287.2023.10520633
- [28] NVIDIA. 2024. NVIDIA DLSS 3. <https://www.nvidia.com/en-us/geforce/technologies/dlss/>
- [29] David Petrescu, Paul A. Warren, Zahra Montazeri, Gabriel Strain, and Steve Pettifer. 2025. The Effects of Visual Behavior and Ego-Movement on Foveated Rendering Performance in Virtual Reality. *SN Computer Science* 6, 4 (4 2025). doi:10.1007/s42979-025-03885-7
- [30] Yanyuan Qin, Shuai Hao, Krishna R. Pattipati, Feng Qian, Subhabrata Sen, Bing Wang, and Chaoyun Yue. 2019. Quality-aware strategies for optimizing ABR video streaming QoE and reducing data usage (MMSys '19). Association for Computing Machinery, New York, NY, USA, 189–200. doi:10.1145/3304109.3306231
- [31] Haseeb Ur Rehman, Shervin Shirmohammadi, Ihab Amer, and Mohamed Hefeeda. 2026. GameSR: Real-Time Super-Resolution for Interactive Gaming. *TechRxiv* preprint. doi:10.36227/techrxiv.177006471.19665190/v1
- [32] Gaurav Sharma, Wencheng Wu, and Edul N Dalal. 2005. The CIEDE2000 Color-Difference Formula: Implementation Notes, Supplementary Test Data, and Mathematical Observations. *Col Res Appl* 30 (2005), 21–30. doi:10.1002/col
- [33] Deniz Ugur, Ihab Amer, and Mohamed Hefeeda. 2025. Decoupling Video Upscaling from Rendering for Cloud Gaming. In *Proceedings of ACM Multimedia Systems Conference (MMSys'25)*, Vol. 1. Stellenbosch, South Africa. doi:10.1145/3712676
- [34] Ritik Vatsal, Shrivatsa Mishra, Rushil Thareja, Mrinmoy Chakrabarty, Ojaswa Sharma, and Jainendra Shukla. 2024. An Analysis of Physiological and Psychological Responses in Virtual Reality and Flat Screen Gaming. *IEEE Transactions on Affective Computing* 15, 3 (7 2024), 1696–1710. doi:10.1109/TAFFC.2024.3368703
- [35] Patricia Vitoria, Lara Raad, and Coloma Ballester. 2020. ChromaGAN: Adversarial Picture Colorization with Semantic Class Distribution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Snowmass Colorado. <https://github.com/pvitoria/ChromaGAN>.
- [36] Zhimin Wang, Xiangyuan Gu, and Feng Lu. 2025. Dominant-Eye-Aware Asymmetric Foveated Rendering for Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics* (7 2025), 1–12. doi:10.1109/tvcg.2025.3593899
- [37] Wireshark Foundation. [n. d.]. Wireshark: Network Protocol Analyzer. <https://www.wireshark.org>
- [38] En Sing Wong, Nur Haliza Abdul Wahab, Faisal Saeed, and Nouf Alharbi. 2022. 360-Degree Video Bandwidth Reduction: Technique and Approaches Comprehensive Review. *Applied Sciences (Switzerland)* 12, 15 (8 2022). doi:10.3390/app12157581
- [39] Wenbo Xu et al. 2023. CVGSR: Cross-View Guided Stereo Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 3193–3203. doi:10.1109/CVPRW59228.2023.00340
- [40] Guanjun Ye, Ge Li, Xiangtao Li, Zhiyong Gao, and Xiaoqi Sun. 2023. Neural Super-Resolution for Omnidirectional Video Streaming. In *Proceedings of the 31st ACM International Conference on Multimedia*. 8123–8131. doi:10.1145/3581783.3612304
- [41] Jiannan Ye, Xiaoxu Meng, Daiyun Guo, Cheng Shang, Haotian Mao, and Xubo Yang. 2024. Neural foveated super-resolution for real-time VR rendering. *Computer Animation and Virtual Worlds* 35, 4 (7 2024). doi:10.1002/cav.2287
- [42] Jooyeol Yun, Sanghyeon Lee, Minhoo Park, and Jaegul Choo. 2023. iColoriT: Towards Propagating Local Hint to the Right Region in Interactive Colorization by Leveraging Vision Transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, Hawaii, 1787–1796. <http://arxiv.org/abs/2207.06831>
- [43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. [n. d.]. *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric*. Technical Report. <https://www.github.com/richzhang/PerceptualSimilarity>.
- [44] Sihao Zhao, Hatem Abou-Zeid, Rami Atavia, Yoga Suhas Kuruba Manjunath, Akram Bin Sediq, and Xiao Ping Zhang. 2021. Virtual Reality Gaming on the Cloud: A Reality Check. In *2021 IEEE Global Communications Conference, GLOBECOM 2021*. Madrid, Spain, 1–6. doi:10.1109/GLOBECOM46510.2021.9685808

- [45] Kaiwen Zheng, Jie Huang, Man Zhou, and Feng Zhao. 2023. *Asymmetric Color Transfer with Consistent Modality Learning*. Technical Report. Vancouver. 2815–2823 pages. <https://github.com/keviner1/SICNet>.
- [46] Chengyi Zou, Shuai Wan, Marc Gorris Blanch, Luka Murn, Marta Mrak, Juil Sock, Fei Yang, and Luis Herranz. 2024. Lightweight Deep Exemplar Colorization via Semantic Attention-Guided Laplacian Pyramid. *IEEE Transactions on Visualization and Computer Graphics* (12 2024), 1–12. doi:10.1109/TVCG.2024.3398791
- [47] Wenbin Zou, Hongxia Gao, Liang Chen, Yunchen Zhang, Mingchao Jiang, Zhongxin Yu, and Ming Tan. 2023. Cross-View Hierarchy Network for Stereo Image Super-Resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Vancouver, 1396–1405. <https://github>.