NIER: Practical Neural-enhanced Low-bitrate Video Conferencing

Anlan Zhang[†], Yuming Hu[‡], Chendong Wang[*], Yu Liu[†], Zejun Zhang[†], Haoyu Gong[‡], Ahmad Hassan[†], Shichang Xu[◇], Zhenhua Li[○], Bo Han[††], Feng Qian[†]
[†]University of Southern California [††]George Mason University [‡]University of Minnesota – Twin Cities
[*]University of Wisconsin – Madison [◇]Google [○]Tsinghua University

Abstract

We present NIER, a video conferencing system that can adaptively maintain a low bitrate (*e.g.,* 10–100 Kbps) with reasonable visual quality while being robust to packet losses. We use key-point-based deep image animation (DIA) as a key building block and address a series of networking and system challenges to make NIER practical. Our evaluations show that NIER significantly outperforms the baseline solutions.

CCS Concepts

• **Information systems** → **Multimedia streaming**.

Keywords

Low-bitrate Video Conferencing, Deep Image Animation, Neural Codecs, Real-time Streaming, Quality-of-Experience

ACM Reference Format:
Anlan Zhang[†], Yuming Hu[‡], Chendong Wang[*], Yu Liu[†], Zejun Zhang[†], Haoyu Gong[‡],, Ahmad Hassan[†], Shichang Xu[◇], Zhenhua Li[○], Bo Han[††], Feng Qian[†]. 2025. NIER: Practical Neural-enhanced Low-bitrate Video Conferencing. In *ACM SIGCOMM 2025 Conference (SIGCOMM '25), September 8–11, 2025, Coimbra, Portugal.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3718958.3750518

1 Introduction

Video conferencing requires a substantial amount of network bandwidth. For example, Zoom requires at least 1.2 Mbps bandwidth for both uplink and downlink in a 1-on-1 video call at 720p [1], which amounts to ~1.05 GB of data for a one-hour session. *Low-bitrate video conferencing* thus benefits multiple stakeholders: streaming platforms spend less on network infrastructures; cellular providers see reduced peak-hour traffic; mobile customers pay less over metered links; and most importantly, end users perceive better quality-of-experience (QoE) under challenging network conditions.

A promising approach to realize low-bitrate video conferencing with decent QoE is to stream low-resolution video frames using traditional codecs such as H264 [12], HEVC [17], and VPX [4], and then apply image enhancement techniques like super-resolution (SR) [16] at the receiver to boost visual quality. One limitation of this approach is that, for efficiency, all traditional codecs incur high temporal dependency; *i.e.,* they produce P-frames whose decoding depends on prior I- or P-frames. As a result, a single packet loss can lead to the undecodability of multiple consecutive frames. Note that packet losses are prevalent in challenging network conditions – a key usage scenario of low-bitrate video streaming.

There are a few techniques to counteract packet losses in real-time video communication, such as retransmissions, forward error correction (FEC) [10, 13], error concealment [19], and loss-resilient neural codecs [7, 9]. However, they suffer from various limitations such as prolonged latency, extra bandwidth cost, low compression efficiency, and high compute overhead, respectively.

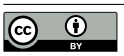
In this work, we develop NIER, a practical low-bitrate video conferencing solution. It can adaptively maintain a low bitrate between 10 and 100 Kbps with reasonable visual quality while being robust to packet losses. Satisfying these design requirements makes NIER suitable for a wide range of usage scenarios, in particular over challenging/metered networks. Under the hood, NIER leverages key-point-based deep image animation (DIA) as a key building block, where the sender transmits sparse key-points alongside a reference image, and the receiver reconstructs the original video frames by animating the reference image using the key-points' motion. To make DIA practical, NIER addresses a series of challenges in networking and system dimensions, including robustly updating reference frames, adapting to fluctuating bandwidth, handling varying packet loss rates, and achieving line-rate frame processing on commodity client devices.

We implement NIER's prototype in 13+ lines of code. Our extensive evaluations (including an IRB-approved user study involving 20 participants) demonstrate that NIER considerably outperforms several baseline solutions (traditional video codecs, super-resolution-enhanced video conferencing [16], forward error coding (FEC) [13], loss-resilient neural codec [7], and naive application of key-point-based DIA) in terms of end-to-end latency, decodable frame ratio, frame rate, video quality, and/or users' quality-of-experience (QoE).

While focused on key-point-based DIA, NIER's high-level design principles are potentially applicable to other neural-based video streaming systems that involve heterogeneous streams (*e.g.,* [7, 16]).

2 Background and Motivation

Deep image animation (DIA) was originally designed to animate a static image using the motion and deformation (*e.g.,* the optical flow [5] between two frames) of a video clip [14, 18, 20]. Recent studies [2, 11] have explored using key-point-based DIA in low-bitrate video calls, where motion and deformation are "encoded" as sparse key-points (consisting of coordinates and attributes) transmitted from the sender. The receiver uses these key-points alongside a high-quality "reference frame" to generate corresponding frames through a pre-trained DIA model. Compared to traditional pixel-based codecs, the key-point representation bears a much lower bitrate. In addition, since each frame is independently encoded into key-points, packet losses affect only corresponding frames, rather



SIGCOMM '25, Coimbra, Portugal

ACM ISBN 979-8-4007-1524-2/2025/09
https://doi.org/10.1145/3718958.3750518

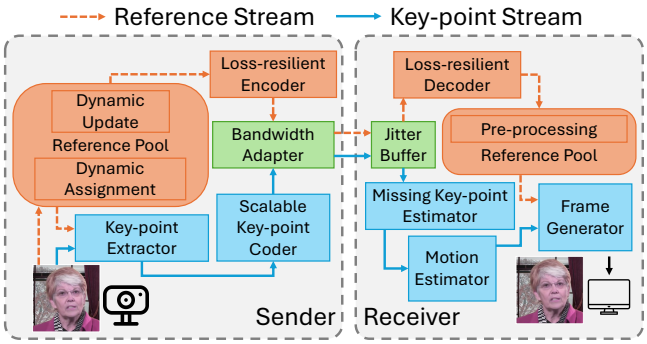


Figure 1: The system architecture of NIER.

than propagating errors across multiple frames. However, existing studies [2, 8, 11] focus on improving key-point-based DIA models while overlooking critical challenges in the networking and system dimensions:

Challenge 1. *When and how to transmit a reference frame?* A reference frame serves a similar role to an I-frame in traditional codecs, as it provides a more recent (and oftentimes better) "baseline" for frame generation. However, key differences between them create unique yet underexplored optimization opportunities – in particular, when and how to transmit a reference frame. In contrast, existing approaches [2, 11] typically send a reference frame only once at the beginning, causing significant quality drop over time.

Challenge 2. *How to adapt DIA to the fluctuating bandwidth?* This can be regarded as "Adaptive Bitrate (ABR) streaming" [6] for DIA, which is largely an uncharted territory.

Challenge 3. *How to handle packet losses?* Likewise, there lack methods allowing DIA to dynamically adapt to varying packet loss rates, limiting its robustness in the real world.

Challenge 4. *How to make DIA practical on commercial off-the-shelf (COTS) devices?* Our measurements show that state-of-the-art key-point-based DIA [15] exhibits poor performance on COTS devices (*e.g.,* 11 FPS with 100+ ms frame processing latency on MacBook Air 2020 [3]), making it falling far short for practical use.

3 System Design of NIER

To the best of our knowledge, NIER is a first practical low-bitrate video conferencing system enhanced by key-point-based DIA. As shown in Figure 1, NIER maintains two streams between the sender and receiver: a key-point stream and a reference stream. It achieves low bitrate by transmitting most video frames as key-points, with adaptive reference frame delivery as needed. We elaborate the design of NIER below.

• To address **Challenge 1**, NIER judiciously updates the reference frame by jointly considering the bandwidth constraint and visual quality impact. A challenge here is that the visual quality groundtruth of a to-be-generated frame is unknown. We thus devise a lightweight approach to predict the visual quality by leveraging a new metric called *self-similarity*, *i.e.,* the similarity between a to-be-generated frame and the reference frame. We find that the self-similarity is highly correlated with the visual quality groundtruth and can be easily derived on the sender side, making it a good predictor. In addition, instead of discarding old reference frames, NIER opportunistically reuses them to further boost the QoE.

• To address **Challenge 2**, our key insight is that the key-point stream and the reference stream, which compete for bandwidth, exhibit distinct characteristics: the key-point stream demands low bandwidth but requires immediate delivery, whereas the reference stream consumes high bandwidth yet remains delay-tolerant. NIER hence employs different strategies for them. For the key-point stream, it adopts a layered encoding scheme that encodes the key-point data into a base layer and three enhancement layers. At runtime, enhancement layers can be flexibly dropped to meet the bandwidth budget. For the reference stream, NIER reshapes its traffic pattern to make it less bursty and more elastic.

• To address **Challenge 3**, We make two observations: (1) similar to pixel-based videos, key-points exhibit temporal locality; and (2) using a recent reference frame with some missing pixels can oftentimes yield a higher generation quality than using an old non-corrupted reference frame. Therefore, for the key-point stream, NIER applies lightweight approaches to infer the missing key-points on the receiver side using historical data when packet loss occurs. For the reference stream, NIER reconstructs a reference frame from partially received segments with negligible overhead.

• To address **Challenge 4**, NIER applies a series of optimizations, including removing redundant computation, pruning/modifying DNN blocks, reducing input data, and pipelining processing stages. Many of these optimizations go beyond standard deep learning inference optimizations by considering the unique characteristics of key-point-based DIA.

4 Implementation and Evaluation

We implement the above components and integrate them into a deployable prototype comprising 13K+ lines of code. We highlight key evaluation results as follows.

• Under ultra-low bandwidth (< 50 Kbps) with a 50 ms one-way delay, NIER achieves 205 (225) ms 50th (95th) percentile (P50 and P95) end-to-end latency, a 99.8% (99.9%) reduction compared to a baseline design where key-point-based DIA is applied in a straightforward manner according to the computer vision literature [2, 11]. In addition, NIER improves the decodable frame ratio by 11.47% under a 10% packet loss rate and improves the visual quality (in PSNR) by 2.03 dB.

• Compared to a SOTA low-bitrate video conferencing solution enhanced by super-resolution [16], NIER improves the P50 (P95) end-to-end latency by up to 98.5% (99.1%), and achieves up to 159x improvement in decodable frame ratio, with a comparable or even better visual quality.

• Compared to a SOTA loss-resilient neural codec [7] and a SOTA FEC scheme [13] for real-time streaming, NIER exhibits much better coding efficiency, in terms of one or more metrics (processing latency, frame rate, and data usage, *etc.*).

• Our IRB-approved user trial involving 20 subjects suggests that NIER outperforms other low-bitrate video conferencing solutions, which uses H264, VP8 and super-resolution [16], by 2.0, 1.45, and 1.7 (in the scale of 1-5), respectively.

Acknowledgments

We thank the anonymous reviewers for their insightful comments.

References

[1] 2025. Zoom system requirements: Windows, macOS, Linux. https://support.zoom.com/hc/en/article?id=zm_kb&sysparm_article=KB0060748.

[2] Madhav Agarwal, Anchit Gupta, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2022. Compressing video calls using synthetic talking heads. *arXiv preprint arXiv:2210.03692* (2022).

[3] Apple Inc. 2020. MacBook with M1 Chip. https://www.apple.com/macbook-air-m1/.

[4] Jim Bankoski, Paul Wilkins, and Yaowu Xu. 2011. Technical overview of VP8, an open source video codec for the web. In *2011 IEEE International Conference on Multimedia and Expo*. IEEE, 1–6.

[5] Steven S. Beauchemin and John L. Barron. 1995. The computation of optical flow. *ACM computing surveys (CSUR)* 27, 3 (1995), 433–466.

[6] Abdelhak Bentaleb, Bayan Taani, Ali C Begen, Christian Timmerer, and Roger Zimmermann. 2018. A survey on bitrate adaptation schemes for streaming media over HTTP. *IEEE Communications Surveys & Tutorials* 21, 1 (2018), 562–585.

[7] Yihua Cheng, Ziyi Zhang, Hanchen Li, Anton Arapin, Yue Zhang, Qizheng Zhang, Yuhan Liu, Kuntai Du, Xu Zhang, Francis Y Yan, et al. 2024. {GRACE}:{Loss-Resilient}{Real-Time} Video through Neural Codecs. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*. 509–531.

[8] Goluck Konuko, Giuseppe Valenzise, and Stéphane Lathuilière. 2021. Ultra-low bitrate video conferencing using deep image animation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4210–4214.

[9] Tianhong Li, Vibhaalakshmi Sivaraman, Pantea Karimi, Lijie Fan, Mohammad Alizadeh, and Dina Katabi. 2023. Reparo: Loss-Resilient Generative Codec for Video Conferencing. *arXiv preprint arXiv:2305.14135* (2023).

[10] Abdelhamid Nafaa, Tarik Taleb, and Liam Murphy. 2008. Forward error correction strategies for media streaming over wireless networks. *IEEE Communications Magazine* 46, 1 (2008), 72–79.

[11] Maxime Oquab, Pierre Stock, Daniel Haziza, Tao Xu, Peizhao Zhang, Onur Celebi, Yana Hasson, Patrick Labatut, Bobo Bose-Kolanu, Thibault Peyronel, et al. 2021. Low bandwidth video-chat compression using deep generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2388–2397.

[12] Iain E Richardson. 2011. *The H. 264 advanced video compression standard*. John Wiley & Sons.

[13] Michael Rudow, Francis Y Yan, Abhishek Kumar, Ganesh Ananthanarayanan, Martin Ellis, and KV Rashmi. 2023. Tambur: Efficient loss recovery for videoconferencing via streaming codes. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. 953–971.

[14] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2377–2386.

[15] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. *Advances in neural information processing systems* 32 (2019).

[16] Vibhaalakshmi Sivaraman, Pantea Karimi, Vedantha Venkatapathy, Mehrdad Khani, Sadjad Fouladi, Mohammad Alizadeh, Frédo Durand, and Vivienne Sze. 2024. Gemino: Practical and robust neural compression for video conferencing. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*. 569–590.

[17] Vivienne Sze, Madhukar Budagavi, and Gary J Sullivan. 2014. High efficiency video coding (HEVC). In *Integrated circuit and systems, algorithms and architectures*. Vol. 39. Springer, 40.

[18] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10039–10049.

[19] Yao Wang and Qin-Fan Zhu. 1998. Error control and concealment for video communication: A review. *Proc. IEEE* 86, 5 (1998), 974–997.

[20] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. 2024. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1481–1490.