AI Video and Audio Enhancement

Mohamed Ahmed Mohamed Sayed
Department of Computer Science
Faculty of Computer and Information Sciences
Ain Shams University
Cairo,Egypt
2021170451@cis.asu.edu.eg

Seif Aldien Ahmed Faheem
Department of Computer Science
Faculty of Computer and Information Sciences
Ain Shams University
Cairo,Egypt
2021170251@cis.asu.edu.eg

Abdelrahman Emad Bayoumy Ali
Department of Computer Science
Faculty of Computer and Information Sciences
Ain Shams University
Cairo,Egypt
2021170300@cis.asu.edu.eg

Nour eldin Mohamed Mounir
Department of Computer Science
Faculty of Computer and Information Sciences
Ain Shams University
Cairo,Egypt
2021170589@cis.asu.edu.eg

Abdullah Ashraf Ahmed Sadek
Department of Computer Science
Faculty of Computer and Information Sciences
Ain Shams University
Cairo,Egypt
2021170317@cis.asu.edu.eg

Seif Ahmed Mohamed Mahmoud
Department of Computer Science
Faculty of Computer and Information Sciences
Ain Shams University
Cairo,Egypt
2021170250@cis.asu.edu.eg

***Abstract*—This paper presents a unified AI-based system for enhancing both video and audio quality using state-of-the-art deep learning models. The project addresses the persistent issue of low-resolution videos and poor audio quality due to factors such as noise, compression artifacts, and motion blur. The proposed solution integrates a Recurrent Video Restoration Transformer (RVRT) for super-resolution and a fine-tuned Demucs model for audio denoising. To overcome hardware limitations, a tile-based processing approach is employed, significantly reducing GPU memory usage without sacrificing quality. Quantitative results show a 1.85 dB improvement in PSNR and 0.0483 increase in SSIM on video benchmarks, while the audio module achieves a PESQ score of 3.15 on the Valentini dataset and 2.91 on VoiceBank+DEMAND. The system is deployed through a web interface that enables real-time or batch enhancement with flexible configuration options. This work contributes a scalable, accessible, and efficient solution for AI-driven media enhancement, with demonstrated potential in content creation, legacy media restoration, and video conferencing. Future work includes integration of advanced transformers, real-time processing, edge deployment, and perceptual quality metrics**

Keywords—Video Enhancement, Audio Denoising, Deep Learning, RVRT, Demucs, Super-Resolution, PESQ, PSNR, SSIM.

I. Introduction

The exponential growth in digital media consumption has heightened expectations for high-quality video and audio content across various domains, including entertainment, education, teleconferencing, and archival restoration. However, many existing video and audio assets suffer from degradation due to low resolution, compression artifacts, background noise, and hardware limitations during acquisition. These quality issues negatively affect user experience and limit the reuse of legacy content.

Recent advances in artificial intelligence (AI), particularly in deep learning, have enabled significant progress in addressing these challenges. Convolutional Neural Networks (CNNs), Transformers, and Recurrent architectures have demonstrated impressive capabilities in video super-resolution, denoising, and texture reconstruction. Similarly, state-of-the-art audio enhancement models have achieved notable success in real-time speech denoising, enabling clearer communication and improved transcription accuracy.

In this work, we propose an integrated AI-based system that enhances both video and audio quality through the combined application of optimized deep learning models. The system leverages the Recurrent Video Restoration Transformer (RVRT) for spatio-temporal video enhancement and the Demucs model for audio noise reduction. A tile-based processing mechanism is introduced to overcome GPU memory limitations, enabling efficient processing on commodity hardware. The system is made accessible through a web-based interface that provides real-time and batch processing capabilities.

This paper presents the methodology, architecture, implementation, and evaluation of the proposed system. Quantitative and qualitative results demonstrate substantial improvements in perceptual and structural quality metrics such as PSNR, SSIM, and PESQ. The system's scalability and usability make it a practical solution for various real-world applications, including media restoration, video conferencing, and content production.

II. Related work

A. **[1]** *EvTexture: Event-driven Texture Enhancement for Video Super-Resolution*

The paper presents a novel approach that leverages high-frequency details from event-based vision to improve texture restoration in Video Super-Resolution (VSR). Named EvTexture, this method introduces a two-branch architecture combining motion learning and a dedicated texture enhancement branch that utilizes event data from neuromorphic cameras. The core innovation is an Iterative Texture Enhancement (ITE) module that refines texture across multiple iterations, extracting temporal details to progressively recover high-quality textures in video frames. The model outperforms state-of-the-art VSR techniques, particularly on texture-rich datasets like Vid4, achieving up to a 4.67 dB gain in PSNR and demonstrating superior perceptual quality. EvTexture is optimized to use fewer parameters and runtime than comparable RGB-based models, performing efficiently on datasets both simulated (Vimeo-90K, REDS) and real-world (CED).

EvTexture's iterative architecture with ConvGRU layers allows it to transfer high-frequency information effectively, leading to smoother and more consistent texture transitions. The model employs a synthetic event voxel grid for texture data representation, which enables it to enhance restoration while maintaining temporal coherence. Its advanced design has a parameter count of 8.9M and a runtime of 136 ms per clip on an NVIDIA V100 GPU, balancing high accuracy with computational efficiency. Additionally, EvTexture+ extends the model by incorporating event-based motion alignment, further improving performance in complex motion scenarios. The paper's comprehensive evaluations highlight EvTexture's practical advantages for high-dynamic-range scenes, reinforcing the value of event-based signals in advancing video super-resolution capabilities.

B. **[2]** *VRT: A Video Restoration Transformer*

The "Video Restoration Transformer (VRT)" introduces a new framework for enhancing video quality by restoring low-quality frames to high-resolution outputs across various tasks, including video super-resolution, denoising, deblurring, frame interpolation, and space-time super-resolution. VRT's design uses a multi-scale approach that processes videos in parallel, enabling efficient handling of long-term dependencies without the limitations of traditional frame-by-frame or sliding window methods. VRT's core components are Temporal Mutual Self-Attention (TMSA) modules and Parallel Warping, which together allow for precise alignment and fusion of features from neighboring frames, efficiently capturing both short and long-range dependencies across multiple resolutions. This approach results in more detailed, high-quality frames with superior performance on several benchmarks, achieving up to 2.16dB higher PSNR compared to state-of-the-art methods.

In terms of specifications, VRT is evaluated on fourteen benchmark datasets and tested on tasks like Vimeo-90K and REDS datasets for video super-resolution. The model demonstrates efficient parameter management and runtime, with 35.6M parameters and a runtime of 243 ms per frame on high-resolution (e.g., 1280x720) inputs. VRT's parallelization capability further supports distributed deployment, making it ideal for large-scale applications. It outperforms recent models, such as BasicVSR++, by integrating both multi-scale feature extraction and enhanced temporal dependencies without recurring performance drops on short or long sequences, establishing a new standard in video restoration performance and efficiency.

C. **[3]** *Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data*

The paper extends the Enhanced Super-Resolution GAN (ESRGAN) to Real-ESRGAN, a model designed to handle real-world image degradation. Real-ESRGAN improves upon previous blind super-resolution models by introducing a high-order degradation model that simulates complex real-world artifacts through multiple degradation processes, including blur, noise, and JPEG compression. Additionally, a U-Net-based discriminator with spectral normalization is used to enhance local texture restoration and stabilize training. This model, trained on synthetic data alone, achieves high-quality image restoration, effectively reducing artifacts like ringing and overshoot, commonly seen in real-world degraded images.

The Real-ESRGAN architecture is efficient with a manageable number of parameters, leveraging a second-order degradation process to balance simplicity and capability. It was trained on datasets like DIV2K, Flickr2K, and OutdoorSceneTraining using NVIDIA V100 GPUs, achieving high perceptual quality across diverse real-world images without requiring real-degradation training data. In experiments, Real-ESRGAN significantly outperformed traditional ESRGAN and other state-of-the-art methods, offering both computational efficiency and superior visual quality, making it well-suited for real-world applications in image enhancement tasks.

D. **[4]** *Recurrent Video Restoration Transformer with Guided Deformable Attention*

The (RVRT) presents a novel model for video restoration tasks such as super-resolution, deblurring, and denoising. Unlike traditional methods that process video frames either in parallel or in sequence, RVRT combines both approaches by processing local neighboring frames in parallel within a globally recurrent structure. This setup achieves a balance between model size, memory usage, and temporal information retention, which is crucial for handling long-range dependencies and reducing noise. The model uses "Guided Deformable Attention" (GDA) for clip-to-clip alignment, which utilizes optical flow to dynamically sample and aggregate multiple relevant locations across frames. This significantly improves efficiency and accuracy in feature alignment, surpassing other alignment techniques by reducing computation while allowing larger receptive fields.

RVRT's structure consists of 10.8 million parameters and achieves top performance on benchmark datasets like REDS and Vimeo-90K with less memory and runtime compared to previous state-of-the-art models. The model runs with a relatively fast runtime of 183 milliseconds per frame and moderate memory use of 1056 MB. It demonstrates substantial gains in PSNR, achieving 32.75 dB on REDS4 with a balanced performance across various video restoration tasks. Training was conducted on datasets such as REDS, Vimeo-90K, DVD, GoPro, and DAVIS, with different setups tailored to video super-resolution and noise levels up to 50. RVRT's guided deformable attention makes it particularly efficient and versatile, with demonstrated success in aligning and enhancing video quality even under significant degradation conditions(RVRT).

E. **[5]** *Real-time Speech Enhancement on Raw Signals with Deep State-space Modeling*

The paper presents a novel approach to online speech enhancement using a deep state-space model (SSM) named aTENNuate. This model is designed to improve real-time denoising of raw speech signals directly, leveraging an efficient autoencoder architecture. Benchmark testing shows it outperforms similar models on the VoiceBank + DEMAND and Microsoft DNS1 datasets. With only 0.84 million parameters and minimal Multiply-Accumulate Operations (MACs), aTENNuate achieves superior Perceptual Evaluation of Speech Quality (PESQ) scores with a latency of 46.5 ms, making it highly suitable for mobile and resource-constrained applications. The model also handles compressed audio inputs at reduced bitrates and sampling frequencies with consistent performance, illustrating its adaptability to low-resource scenarios.

The architecture of aTENNuate relies on SSM layers that capture long-range temporal relationships in speech, enhancing the model's ability to remove noise effectively. These SSM layers operate with stable linear recurrent units, enabling efficient real-time inference. To maintain causality, the model excludes bidirectional layers and pre/post-processing, keeping raw waveform inputs intact. Trained on the VCTK and LibriVox datasets, the model employs a combination of SmoothL1 and spectral losses for optimization, achieving a high PESQ score of 3.27 on VoiceBank + DEMAND. aTENNuate's lightweight, real-time performance and its capacity to enhance quality in noisy and low-quality audio environments position it as an advancement over prior speech enhancement models.

F. **[6]** *Audio Enhancement for Computer Audition– An Iterative Training Paradigm Using Sample Importance*

The paper introduces a robust framework for audio enhancement (AE) to improve performance across computer audition tasks like automatic speech recognition (ASR), speech command recognition (SCR), speech emotion recognition (SER), and acoustic scene classification (ASC). Unlike traditional models that treat audio enhancement separately, this framework jointly trains the AE and computer audition (CAT) models using iterative optimization and sample importance, allowing the AE module to prioritize difficult samples that negatively affect CAT performance. Evaluated on noisy datasets, this method demonstrates substantial improvements in accuracy across tasks, especially at low signal-to-noise ratios (SNRs), indicating improved resilience to background noise. Models based on a U-Net architecture operate in the frequency domain, using STFT to filter noise from the spectrogram before reconstruction.

The framework employs various training paradigms to benchmark effectiveness, including cold cascade, data augmentation, and multi-task learning, with iterative optimization consistently achieving the best results by dynamically adapting AE output to CAT demands. Using datasets like LibriSpeech, DEMoS, and AudioSet, the authors report that this iterative approach outperforms state-of-the-art models (like DFNet-3 and MetricGAN+) in noisy environments. Evaluation metrics show an average accuracy increase across tasks by focusing on challenging samples. The study concludes that AE methods can be significantly optimized by integrating sample importance and iteratively aligning them with CAT models, paving the way for broader applications in real-life noisy conditions.

G. **[7]** *Noise-aware Speech Enhancement using Diffusion Probabilistic Model*

The paper introduces the Noise-aware Speech Enhancement (NASE) approach, which leverages a diffusion probabilistic model to enhance noisy speech signals by focusing on noise-specific features. The proposed model uses a noise classification (NC) module to generate an acoustic embedding, acting as a "noise conditioner" during the reverse diffusion process to improve noise reduction. A multi-task learning setup jointly optimizes the speech enhancement (SE) and noise classification tasks, refining the noise conditioner's specificity. Tested on the VoiceBank-DEMAND dataset, NASE demonstrated significant improvements in perceptual evaluation of speech quality (PESQ), extended short-time objective intelligibility (ESTOI), and scale-invariant signal-to-distortion ratio (SI-SDR), especially when handling unseen noise types. NASE functions as a plug-and-play addition to various diffusion SE models, including CDiffuSE, StoRM, and SGMSE+, achieving improvements in PESQ scores, with the highest gains seen on SGMSE+.

The model architecture incorporates a BEATs pre-trained model for enhanced noise classification, featuring 12 Transformer encoder layers and 768 embedding units. Key experiments demonstrated that NASE outperformed conventional models across multiple unseen noise levels and types (e.g., Helicopter, Baby-cry, Crowd-party), achieving a maximum PESQ score of 3.01 on SGMSE+ and up to 61% improvement on intelligibility metrics in high noise conditions. NASE's efficiency in integrating noise-specific information through various conditioning methods, like addition and concatenation, shows promise for real-world noisy environments and reinforces the approach's generalizability across diffusion-based SE backbones.

III. SYSTEM ARCHITECTURE

Figure 1 illustrates the workflow of the system, which is designed to enhance the video and audio quality.

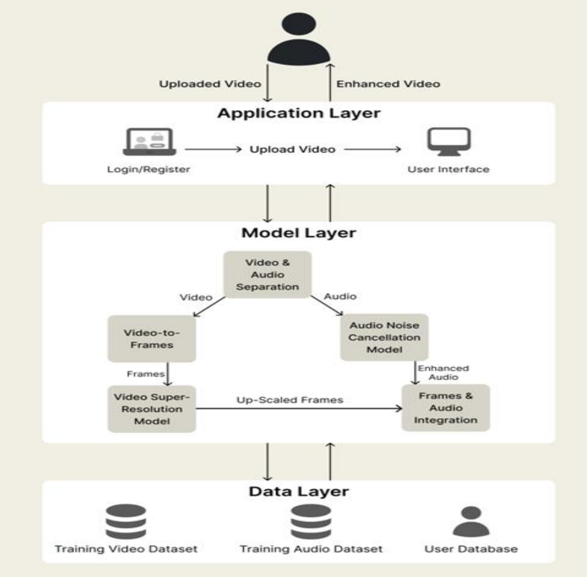


Figure 1 System Architecture

The proposed AI-based video and audio enhancement system is designed as a modular, scalable, and efficient architecture that integrates state-of-the-art models for media quality improvement. The system is structured into three primary layers: the Application Layer, Model Layer, and Data Layer. Each layer is responsible for distinct aspects of the system's functionality, enabling modular development, maintainability, and performance optimization.

A. Application Layer

The Application Layer serves as the user-facing interface of the system, facilitating interaction and control over the enhancement process. It is implemented as a web-based interface that allows users to upload video files, configure enhancement parameters, and retrieve enhanced outputs. Key components of this layer include:

User Interface (UI): A responsive web interface enabling users to interact with the system, select processing modes, and monitor progress.

Authentication Module: A login and registration system for user account management and session tracking.

File Management System: Manages file uploads, processing queues, and retrieval of enhanced media.

This layer ensures ease of use and accessibility for non-technical users while supporting real-time interaction with backend processing modules.

B. Model Layer

The Model Layer constitutes the core processing engine of the system, where advanced deep learning models perform enhancement tasks independently for video and audio components. It comprises the following submodules:

Media Separation Module: Automatically extracts video frames and audio signals from uploaded video files to allow separate enhancement processing.

Video Enhancement Module: Utilizes the Recurrent Video Restoration Transformer (RVRT) to enhance video quality. The model is optimized with residual blocks and tile-based processing for improved performance on limited hardware.

Audio Enhancement Module: Employs a fine-tuned Demucs model to perform speech denoising and audio enhancement in both time and frequency domains.

Media Recombination Module: Merges the enhanced video frames and denoised audio into a single synchronized output video.

The layer is implemented with modularity and extensibility in mind, allowing future replacement or addition of models with minimal architectural changes.

C. Data Layer

The Data Layer handles persistent storage and retrieval of user data, model training datasets, and processing history. It consists of:

Training Datasets: Includes the REDS dataset for video super-resolution and the Valentini and VoiceBank+DEMAND datasets for audio enhancement training.

User Database: Maintains user profiles, authentication tokens, and previous processing logs.

Media Storage: Temporarily stores uploaded, processed, and enhanced media files during system operation.

This layer ensures data integrity, reproducibility of enhancement operations, and traceability of user activity.

D. Design Considerations

To accommodate the computational demands of high-resolution video processing, the system adopts several optimization strategies, including:

Tile-Based Processing: Splits video frames into smaller segments to reduce GPU memory consumption, enabling efficient processing even on mid-range hardware.

Residual Connections: Incorporated into the RVRT pipeline to improve gradient flow and enhance fine detail preservation in video outputs.

Model Fine-Tuning: Transfer learning techniques are applied to adapt pre-trained models to the specific enhancement tasks and noise profiles encountered in real-world scenarios.

This architectural design enables the system to achieve high enhancement quality while maintaining practical execution times and hardware requirements.

IV. PROPOSED METHODLOGY

This section details the methodologies employed for enhancing both video and audio streams using deep learning techniques. The proposed system integrates advanced models and optimization strategies to achieve high-quality media enhancement while maintaining computational efficiency. The methodology is divided into two main pipelines: video enhancement and audio enhancement.

A. Video Enhancement Methodology

Figure 2 illustrates the pipeline of the model, which is designed to enhance the video quality.

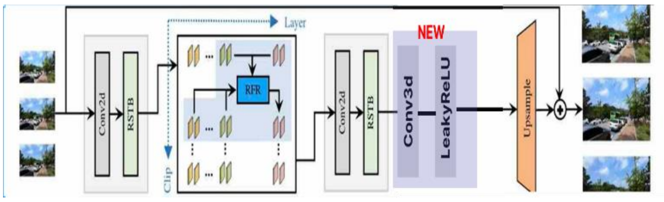


Figure 2 Video Model Architecture

The video enhancement pipeline leverages the Recurrent Video Restoration Transformer (RVRT), which combines recurrent and parallel processing to handle long-range temporal dependencies across frames. The following methods are utilized to optimize performance and quality:

1. RVRT-Based Enhancement

The RVRT model is adapted to process degraded video sequences by capturing spatial and temporal features using guided deformable attention. It processes neighboring frames in parallel within a recurrent framework, balancing quality and computational cost.

2. Tile-Based Processing

To address memory constraints on standard hardware, a tile-based strategy is employed. Each video frame is divided into smaller non-overlapping tiles that are processed independently. This reduces GPU memory consumption by up to 50% without sacrificing output quality. After processing, tiles are reassembled to reconstruct the full frame.

3. Residual Block Enhancement

To further improve detail preservation, additional residual blocks are introduced before the upsampling layers in the RVRT model. These blocks consist of stacked 3D convolutional layers followed by LeakyReLU activations and skip connections, which help preserve fine-grained textures and mitigate vanishing gradient problems during training.

B. Audio Enhancement Methodology

The audio enhancement pipeline is based on a modified version of the Demucs model, originally developed for source separation, and adapted here for speech denoising.

1. Demucs Model Integration

The Demucs architecture adopts a U-Net structure that operates in both time and frequency domains. It includes convolutional encoder and decoder blocks with skip connections that enable precise reconstruction of clean speech from noisy inputs.

2. Transfer Learning and Fine-Tuning

The pre-trained Demucs model is fine-tuned using domain-specific datasets such as Valentini-noise and VoiceBank+DEMAND. Transfer learning allows the model to adapt to varied noise profiles and recording conditions. Multi-dataset training ensures robustness across diverse environmental noises and speech patterns.

3. Noise Suppression and Intelligibility Preservation

To ensure high speech intelligibility, the model applies spectral and waveform-level loss functions during training. These include a combination of SmoothL1 loss and perceptual metrics to maintain naturalness while suppressing background noise.

C. Post-Processing and Media Integration

After independent processing, the enhanced video frames and denoised audio stream are recombined. Synchronization is preserved through frame-to-audio alignment, ensuring that the final output maintains both visual and auditory coherence. The final video is encoded and made available for download or streaming through the web interface

D. Evaluation Metrics

To assess the effectiveness of the proposed methodologies, the following metrics are used:

1. Video Quality: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM)
2. Audio Quality: Perceptual Evaluation of Speech Quality (PESQ)

These objective metrics guide model optimization and serve as benchmarks for comparing against state-of-the-art approaches.

V. Dataset and Results

A. Datasets

Video: The REDS dataset was used for training the video super-resolution model, while the Vid4 dataset served as the primary benchmark for testing.

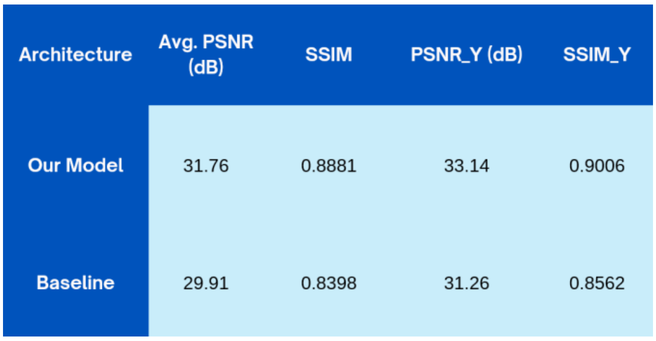
Audio: The Valentini-noise dataset and the VoiceBank+DEMAND corpus were used for training and testing the speech denoising model.

These datasets provide diverse and realistic conditions for evaluating the robustness of the proposed system.

B. Experimental Results

1. Video Enhancement

Figure 3 illustrates the qualitative evaluation metrics for the video models



Architecture	Avg. PSNR (dB)	SSIM	PSNR_Y (dB)	SSIM_Y
Our Model	31.76	0.8881	33.14	0.9006
Baseline	29.91	0.8398	31.26	0.8562

Figure 3 Video Model Results

Quantitative evaluation shows that the enhanced RVRT model outperforms baseline implementations. On the REDS benchmark:

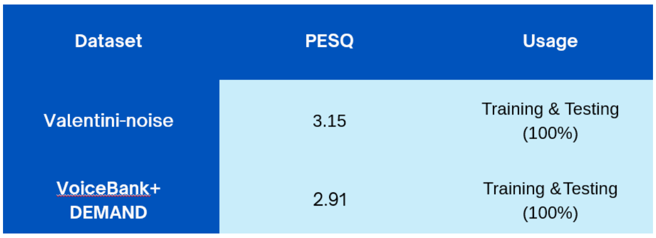
Average PSNR: Improved by 1.85 dB compared to the baseline.

SSIM: Increased by an average of 0.0483, indicating better structural similarity and texture preservation.

Efficiency: Tile-based processing reduced GPU memory usage by approximately 50%, enabling the handling of high-resolution frames on mid-range GPUs without significant degradation in processing speed.

2. Audio Enhancement

Figure 4 illustrates the qualitative evaluation metric for the audio model across different datasets



Dataset	PESQ	Usage
Valentini-noise	3.15	Training & Testing (100%)
VoiceBank+ DEMAND	2.91	Training &Testing (100%)

Figure 4 AudioModel Results

The fine-tuned Demucs model demonstrated effective noise reduction while maintaining speech clarity:

PESQ Score: Achieved 3.15 on the Valentini-noise dataset and 2.91 on VoiceBank+DEMAND, reflecting significant improvements in perceptual speech quality.

Generalization: Robust performance was observed across varying noise levels and unseen test scenarios.

3. Integrated Performance

The combined system successfully produces synchronized, high-quality video and audio output. End-to-end processing time scales linearly with video length and tile size, offering practical batch or near real-time processing modes.

VI. CONCLUSION

This paper presented an integrated AI-based system for video and audio enhancement that addresses common quality degradation issues in digital media. By combining a Recurrent Video Restoration Transformer (RVRT) for video super-resolution with a fine-tuned Demucs model for speech denoising, the proposed solution delivers substantial improvements in both visual and auditory quality.

Experimental results demonstrate that the system achieves significant gains in objective quality metrics, including an average PSNR improvement of 1.85 dB and an SSIM increase of 0.0483 for video sequences, as well as PESQ scores of up to 3.15 for audio enhancement. The introduction of tile-based processing optimizes GPU memory usage, making the system feasible for deployment on standard hardware configurations without compromising performance.

Unlike prior solutions that address video or audio enhancement separately, the proposed system synchronizes both processes within a unified pipeline, ensuring consistent quality improvement across both modalities. The user-friendly web interface further enhances accessibility, enabling real-time or batch processing for a wide range of practical applications, including legacy media restoration, video conferencing, and online content creation.

Future work will focus on extending the system's capabilities by exploring advanced transformer architectures, cross-modal learning strategies, and real-time processing optimizations. Additionally, deployment on edge devices and mobile platforms will be investigated to broaden the system's usability. Further research will also consider integrating perceptual quality metrics and adaptive content-aware enhancement to align more closely with user preferences and diverse media scenarios.

The promising results confirm that the proposed AI-based video and audio enhancement system provides a practical and scalable solution for improving digital media quality, with significant potential for real-world impact across multiple domains.

REFERENCES

[1] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. C. Loy, "EvTexture: Event-driven Texture Enhancement for Video Super-Resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5421–5437, Aug. 2024.

[2] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "VRT: A Video Restoration Transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 17482–17492.

[3] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 1905–1914.

[4] J. Liang, Y. Fan, X. Xiang, R. Ranjan, E. Ilg, S. Green, et al., "Recurrent Video Restoration Transformer with Guided Deformable Attention," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 378–393, Dec. 2022.

[5] M. Defossez, G. Synnaeve, and E. Zeghidour, "Real-Time Speech Enhancement on Raw Signals with Deep State-Space Modeling," in *Proc. Interspeech*, Dublin, Ireland, Aug. 2023, pp. 1354–1358.

[6] Y. Chen, X. Li, Z. Huang, and S. Watanabe, "Audio Enhancement for Computer Audition: An Iterative Training Paradigm Using Sample Importance," in *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1102–1114, Mar. 2023.

[7] H. Zhang, T. Li, J. Xu, Y. Qian, and Z. Tan, "Noise-Aware Speech Enhancement Using a Diffusion Probabilistic Model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 2023, pp. 116–120.