

Article

A3DSimVP: Enhancing SimVP-v2 with Audio and 3D Convolution

Junfeng Yang ^{1,2}, Mingrui Long ¹, Hongjia Zhu ^{1,*}, Limei Liu ^{1,2,*}, Wenzhi Cao ^{1,2}, Qin Li ^{1,2} and Han Peng ^{1,2} 

¹ School of Intelligent Robotics, Hunan University of Technology and Business, Changsha 410205, China; b12100031@hnu.edu.cn (J.Y.); 230120854010@stu.hutb.edu.cn (M.L.); wenzcao@hutb.edu.cn (W.C.); qinli@hutb.edu.cn (Q.L.); han.peng@hutb.edu.cn (H.P.)

² Xiangjiang Laboratory, Changsha 410205, China

* Correspondence: 250220854157@stu.hutb.edu.cn (H.Z.); seagullm@hutb.edu.cn (L.L.)

Abstract

In modern high-demand applications, such as real-time video communication, cloud gaming, and high-definition live streaming, achieving both superior transmission speed and high visual fidelity is paramount. However, unstable networks and packet loss remain major bottlenecks, making accurate and low-latency video error concealment a critical challenge. Traditional error control strategies, such as Forward Error Correction (FEC) and Automatic Repeat Request (ARQ), often introduce excessive latency or bandwidth overhead. Meanwhile, receiver-side concealment methods struggle under high motion or significant packet loss, motivating the exploration of predictive models. SimVP-v2, with its efficient convolutional architecture and Gated Spatiotemporal Attention (GSTA) mechanism, provides a strong baseline by reducing complexity and achieving competitive prediction performance. Despite its merits, SimVP-v2's reliance on 2D convolutions for implicit temporal aggregation limits its capacity to capture complex motion trajectories and long-term dependencies. This often results in artifacts such as motion blur, detail loss, and accumulated errors. Furthermore, its single-modality design ignores the complementary contextual cues embedded in the audio stream. To overcome these issues, we propose A3DSimVP (Audio- and 3D-Enhanced SimVP-v2), which integrates explicit spatio-temporal modeling with multimodal feature fusion. Architecturally, we replace the 2D depthwise separable convolutions within the GSTA module with their 3D counterparts, introducing a redesigned GSTA-3D module that significantly improves motion coherence across frames. Additionally, an efficient audio–visual fusion strategy supplements visual features with contextual audio guidance, thereby enhancing the model's robustness and perceptual realism. We validate the effectiveness of A3DSimVP's improvements through extensive experiments on the KTH dataset. Our model achieves a PSNR of 27.35 dB, surpassing the 27.04 of the SimVP-v2 baseline. Concurrently, our improved A3DSimVP model reduces the loss metrics on the KTH dataset, achieving an MSE of 43.82 and an MAE of 385.73, both lower than the baseline. Crucially, our LPIPS metric is substantially lowered to 0.22. These data tangibly confirm that A3DSimVP significantly enhances both structural fidelity and perceptual quality while maintaining high predictive accuracy. Notably, A3DSimVP attains faster inference speeds than the baseline with only a marginal increase in computational overhead. These results establish A3DSimVP as an efficient and robust solution for latency-critical video applications.



Academic Editor: Maciej Ławryńczuk

Received: 30 September 2025

Revised: 19 December 2025

Accepted: 23 December 2025

Published: 25 December 2025

Copyright: © 2025 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Keywords: video prediction; multimodal fusion; spatiotemporal modeling; 3D convolution

1. Introduction

Real-time video streaming plays an increasingly important role in domains such as live broadcasting, cloud gaming, remote healthcare, and defense [1]. Low-latency transmission is critical for enabling immersive interactions in live streaming. This requirement is particularly stringent for mobile users, who demand high performance despite limited resources [2]. Due to high compression rates and error propagation, video streams remain vulnerable to packet loss and bit errors. Traditional error control mechanisms, such as Forward Error Correction (FEC) [3,4], Automatic Retransmission Request (ARQ) [5], and Error Concealment (EC) [6,7], have made significant progress but face limitations in real-time contexts. Although FEC and ARQ improve reliability, they inevitably incur bandwidth overhead and additional latency, conflicting with strict low-latency requirements. Similarly, receiver-side EC methods often struggle under high-packet-loss or rapid-motion conditions, resulting in artifacts such as blurring and texture loss. To address these challenges, video prediction has emerged as a compelling alternative. By leveraging spatiotemporal correlations to predict subsequent frames, it effectively conceals errors and preserves video quality without the latency penalties of traditional retransmission schemes [7]. Table 1 summarizes the characteristics of these approaches.

Table 1. Comparison of mainstream video prediction algorithms in terms of advantages, disadvantages, and application scenarios.

Method	Advantages	Disadvantages	Typical Application Scenarios
Diffusion Model [8–13]	Generates high-quality images	Training and reasoning extremely slow	High-quality future frame generation AI generates images or videos
	Modelable uncertainty Supports multimodal control	Poor inter-frame consistency High computational cost	
Optical Flow [14–21]	Great consistency in movement	Difficult to predict large changes in optical flow	Short term prediction of pedestrians or traffic in motion
	Physically interpretable	Accumulated errors are large and computational complexity is high Unable to solve the prediction of lighting changes in scenes	object tracking
Interpolation [22–27]	Time series smoothing Transitional nature	Need keyframes Unable to predict new content	Video frame insertion Slow-motion generation
	Supplementary frames	Need to refer to forward frames (future frames cannot be obtained in the video streaming field)	

In this domain, deep learning-based approaches, particularly convolutional neural networks (CNNs), have become the mainstream solutions. Among them, the SimVP-v2 model [28] provides an outstanding benchmark. Traditional models often rely on complex recurrent structures (e.g., PredRNN [29]), large encoder–decoder frameworks (e.g., UNet [30]), or hybrid architectures combining CNNs with Transformers (e.g., FCIHMRT [31]). In contrast, SimVP-v2 innovatively adopts a streamlined, fully convolutional architecture. This model demonstrates that advanced spatiotemporal predictive

performance can be achieved by stacking standard convolutional layers and incorporating an efficient Gated Spatiotemporal Attention (GSTA) [32] mechanism. Such a concise design significantly reduces floating-point operations (FLOPs), delivers high inference throughput and efficiency, and achieves an excellent balance between performance and computational cost, establishing it as a suitable baseline for latency-critical applications. Despite its efficiency breakthroughs and strong performance across multiple datasets, deeper analysis reveals that SimVP-v2 still suffers from critical limitations, restricting its applicability in scenarios demanding high fidelity and extreme robustness.

However, we identify two primary limitations in the SimVP-v2 baseline. First, its reliance on 2D convolutions restricts temporal modeling capacity. This implicit approach proves inadequate for complex scenes, often resulting in severe edge blurring and texture loss during long-term prediction. Second, the exclusive focus on visual input neglects crucial acoustic cues. Audio signals offer strict temporal alignment and semantic context, where acoustic variations can effectively predict visual transitions. To address these challenges, we propose A3DSimVP (Audio- and 3D-Enhanced SimVP-v2). We replace standard 2D operators with explicit 3D spatiotemporal convolutions to capture fine-grained motion dynamics. Furthermore, we incorporate a lightweight audio–visual fusion mechanism to leverage cross-modal constraints. This architecture maintains the computational efficiency of SimVP-v2 while achieving superior temporal coherence and high-fidelity reconstruction. To be specific, the main contributions of this work are summarized as follows:

- We propose the A3DSimVP framework, which achieves explicit spatiotemporal modeling by replacing 2D operators with 3D depthwise separable convolutions. This design captures coherent motion trajectories and long-term dependencies directly. As a result, it effectively mitigates motion blur and detail loss.
- We introduce an efficient audio–visual fusion strategy that integrates synchronized audio features as auxiliary context. By leveraging lightweight channel attention and gating mechanisms, we enhance prediction robustness. This significantly improves performance when visual information is ambiguous.
- Extensive evaluations demonstrate that A3DSimVP achieves an optimal balance between high-fidelity prediction and deployment efficiency. On the UCF101 benchmark, our model outperforms the SimVP-v2 baseline, reducing Mean Squared Error (MSE) [33] by 9.7% and Mean Absolute Error (MAE) [33] by 11.9%. While maintaining predictive performance comparable to, or slightly exceeding, the SimVP-v2 baseline, our architecture offers a decisive advantage in inference speed. Specifically, analysis on the Motion-X dataset reveals that A3DSimVP incurs a moderate 15.0% increase in FLOP consumption but yields a remarkable 63.6% gain in inference throughput, raising the speed from 12.75 to 20.86 samples/s. This trade-off confirms that our model effectively utilizes parallel computing resources to surpass the baseline’s efficiency, making it highly suitable for latency-critical applications.

2. Related Works

The rapid evolution of multimedia communications and streaming applications has established efficient video transmission and recovery under constrained bandwidth and unstable networks as a critical challenge in both computer vision and communications [34]. To address this, researchers have proposed diverse approaches, spanning from traditional error control and retransmission strategies to advanced video prediction and multimodal learning methods. While these studies establish a foundational framework for mitigating packet loss and distortion, significant limitations persist, necessitating the development of more robust and intelligent solutions.

2.1. Traditional Transmission and Error Control Methods

Efficient video transmission over unstable, bandwidth-constrained networks remains a critical challenge. Early research prioritized transmission robustness through Forward Error Correction (FEC) [3,4] and Automatic Repeat reQuest (ARQ) [35]. Although FEC enhances reliability by introducing redundant bits for reconstruction, it inevitably imposes bandwidth overhead and latency penalties that conflict with real-time constraints. Similarly, ARQ relies on feedback mechanisms for retransmission; while effective for point-to-point communication, it proves inefficient in multicast environments. In contrast, Error Concealment (EC) [6,7] operates at the receiver to recover corrupted frames by leveraging spatial and temporal redundancies, such as motion compensation [23]. However, EC capability degrades significantly under severe packet loss or rapid motion, often resulting in visual artifacts and error accumulation. Consequently, these traditional methods face intrinsic limitations regarding bandwidth, latency, and recovery capability, necessitating the exploration of advanced predictive solutions.

2.2. Video Prediction Method Based on Deep Learning or Optical Flow

To overcome the limitations of traditional transmission methods, recent research has focused on designing neural architectures capable of effectively capturing the complex spatiotemporal dynamics of video [36,37]. These deep learning-based approaches can be categorized into five distinct paradigms:

- **RNN-based Methods:** Early deep learning models, such as PredRNN [29], utilize recurrent structures and spatiotemporal long short-term memory (ST-LSTM [38]) units to model temporal dynamics sequentially. While effective for short-term prediction, their reliance on step-by-step processing often leads to high computational costs and error accumulation in long-term forecasting, limiting their suitability for low-latency streaming.
- **CNN-based Methods:** To address the efficiency bottlenecks of recurrent networks, CNN-based models like SimVP [39] and SimVP-v2 [28] have emerged as efficient alternatives. SimVP-v2 replaces complex recurrent units with a pure convolutional architecture employing a Gated Spatiotemporal Attention (GSTA) [32] mechanism. This concise design significantly reduces floating-point operations (FLOPs) and improves inference throughput, establishing a strong baseline for latency-sensitive tasks.
- **Transformer-based Approaches:** Architectures such as DPT [40] and TAU [41] leverage self-attention mechanisms to capture global long-range dependencies across video frames. Beyond pure transformer designs, hybrid frameworks like FCIHMRT [31] integrate local convolutional features (e.g., via Res2Net) with global transformer contexts through cross-layer interaction mechanisms. Although these methods improve feature representation and temporal consistency compared to local convolutions, the quadratic complexity associated with self-attention often incurs significant computational overhead, posing challenges for deployment on resource-constrained edge devices.
- **Generative Diffusion Models:** Probabilistic approaches, including RaMVD [8], employ stochastic differential processes and 3D convolutions to generate high-fidelity video content. While diffusion models achieve superior visual quality and detail preservation compared to deterministic methods, their reliance on iterative sampling results in slow inference speeds, rendering them impractical for real-time recovery.
- **Motion-Guided Methods:** Recognizing the importance of temporal coherence, motion-guided approaches integrate explicit motion modeling as a prior. Modern optical flow estimators like RAFT [14] and PWC-Net [21] provide high-precision pixel-level motion but are computationally intensive. Consequently, recent works such as VideoFlow [16] and PhyCoPredictor [18] embed these flow priors directly into generative frameworks.

This fusion combines the physical plausibility of flow with the visual fidelity of generative models [18], though optimizing the computational cost of such hybrid systems remains a critical challenge.

In summary, while existing architectures establish a solid foundation for video prediction, they consistently face a trade-off between computational efficiency and generative quality. More critically, these methods predominantly operate within a single visual modality, thereby neglecting the synchronized audio signals that offer vital semantic and temporal context. This oversight limits their robustness in complex, dynamic scenes, motivating the need to investigate multimodal paradigms that can leverage cross-modal cues for enhanced prediction stability.

3. Enhancing SimVP-v2 with Audio and 3D Convolution

This section details the proposed A3DSimVP framework, designed to address the limitations of SimVP-v2 regarding implicit temporal modeling and single-modality constraints. It introduces explicit spatiotemporal representation learning via 3D convolutions and integrates audio–visual fusion through a lightweight mechanism [42,43]. Prioritizing the balance between efficiency and predictive fidelity, the architecture provides a robust solution for real-time video stream recovery.

3.1. Overall Framework and Problem Formulation

The A3DSimVP framework addresses video blurriness and insufficient temporal dependency modeling, limitations often inherent in conventional spatiotemporal prediction. Building upon the fully convolutional architecture of SimVP-v2, we enhance the framework by integrating explicit 3D spatiotemporal modeling and multimodal information fusion.

For the spatiotemporal prediction task, given a video sequence of T historical frames $V_{in} = \{v_k\}_{t-T+1}^t$ and the corresponding audio feature sequence A , the objective is to predict the subsequent T' future frames V_{out} . Each individual frame v_k has dimensions $R^{C \times H \times W}$. Our model M_Φ defines a mapping from multimodal inputs to the future video sequence, expressed as:

$$M_\Phi : (V_{in}, A) \mapsto V_{out} \quad (1)$$

where M_Φ denotes the complete model parameterized by Φ . The training objective is to identify the optimal parameters Φ^* that minimize the prediction error L . Following standard practice, we employ the MSE as the loss function:

$$\Phi^* = \arg \min_{\Phi} L(M_\Phi(V_{in}, A), V_{out}) \quad (2)$$

The complete prediction process of A3DSimVP model is shown in Figure 1.

As depicted in Figure 1, the A3DSimVP model is built upon an Encoder–Predictor–Decoder structure. The Encoder is responsible for extracting spatiotemporal features from historical frames. Subsequently, the Predictor employs an Attention-based 3D module to forecast the evolution of these features within the latent space. Finally, the Decoder reconstructs the future video frames based on the predicted latent features.

In addition to the visual stream, our framework incorporates an audio-processing pathway to enhance multimodal prediction. As illustrated in Figure 1, the audio features are first passed through a squeeze-and-excitation (SE) block to refine their representations. These refined audio embeddings are then fused with the visual features, forming a joint representation that is fed into the Encoder. This multimodal fusion enables the model to capture cross-domain correlations between motion cues in video and semantic information embedded in audio signals.

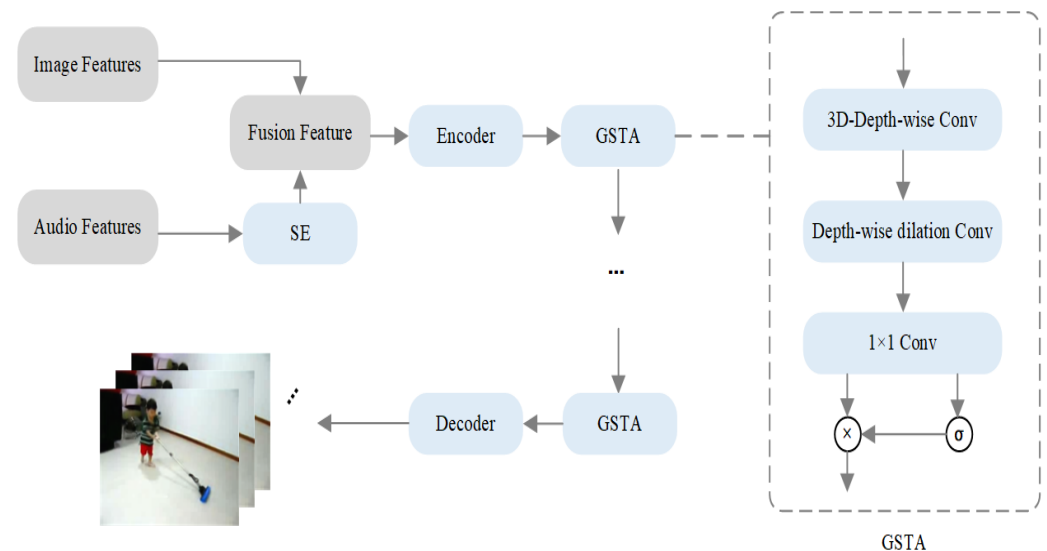


Figure 1. Schematic illustration of the A3DSimVP prediction framework.

Formally, let $A \in \mathbb{R}^{T \times d_a}$ denote the sequence of extracted audio features, and $V_{in} \in \mathbb{R}^{T \times C \times H \times W}$ denote the input video frames. The fusion process can be expressed as:

$$F = \text{Fusion}(\text{SE}(A), V_{in}) \quad (3)$$

where $\text{SE}(\cdot)$ represents the squeeze-and-excitation operation applied to the audio features, and $\text{Fusion}(\cdot)$ denotes the feature combination step. The resulting multimodal feature F is then propagated through the Encoder–Predictor–Decoder pipeline, thereby enabling A3DSimVP to exploit complementary information across modalities.

3.2. Explicit Spatiotemporal Enhancement: 3D Convolution and GSTA Module

Compared with the baseline model SimVP-v2, which relies on 2D convolution (Conv2d) and aggregates temporal information implicitly through 1×1 channel-wise operations, the proposed A3DSimVP introduces a fundamental architectural enhancement by employing 3D convolution (Conv3d).

The SimVP-v2 model leverages 2D convolution for explicit spatial feature extraction across the Height (H) and Width (W) dimensions of each frame. However, it treats the Time Sequence Length (T) implicitly, relying on the Gated Spatiotemporal Attention (GSTA) module to infer inter-frame motion later in the pipeline. This two-stage, implicit approach often leads to accumulated errors and motion blur.

In contrast, our A3DSimVP model utilizes 3D convolution, which is designed to perform explicit spatiotemporal feature extraction. The 3D kernel operates simultaneously across all three dimensions (T, H, and W), allowing our model to directly capture local motion dynamics along with spatial appearance information from the onset. This unified representation is then fed into the GSTA module, enabling the attention mechanism to focus on high-level, long-range dependencies rather than basic motion recovery. Consequently, our architecture significantly improves the structural integrity and perceptual quality of the predicted videos.

This modification enables explicit spatiotemporal feature extraction, allowing the model to jointly learn spatial and temporal representations in a unified manner. As a result, the architecture achieves a more comprehensive understanding of frame-to-frame motion dynamics and structural variations, ultimately improving the accuracy of predictions and the perceptual quality of the reconstructed videos.

As illustrated in Figure 2, there is a fundamental difference in how temporal information is processed. The SimVP-v2 (Figure 2b) utilizes 2D convolution, which extracts explicit spatial features ($H \times W$) while treating the time sequence (T) implicitly. In contrast, our A3DSimVP (Figure 2a) employs explicit 3D depthwise convolution, operating across the entire spatiotemporal volume ($T \times H \times W$) to directly model inter-frame motion. Here, T , H , and W represent the time sequence length, image height, and image width, respectively. This architectural advantage effectively mitigates issues like blurry edges and motion inconsistency, leading to a significant enhancement in the structural and perceptual fidelity of the predicted frames.

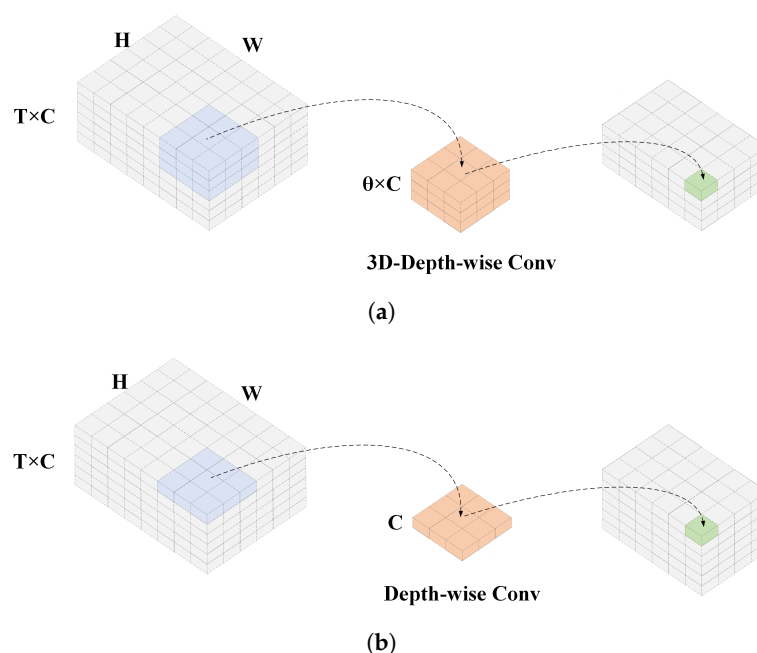


Figure 2. Comparison of feature extraction mechanisms between 3D and 2D convolutions in temporal modeling. The light blue regions indicate the input receptive fields, orange blocks represent the convolutional kernels, and green blocks signify the extracted output features. (a) Visualization of 3D Convolution Feature; (b) Visualization of 2D Convolution Feature.

Within the Translator T , we design an enhanced Gated Spatiotemporal 3D module (GSTA-3D), which achieves explicit temporal modeling via a structured 3D convolution group K (see Figure 2a, in contrast to the implicit modeling of SimVP-v2 in Figure 2b). Here, K denotes the size of the three-dimensional convolutional kernel along the temporal dimension. To provide a clearer understanding of this improvement, the structural diagram of the proposed GSTA module is illustrated below.

Consequently, by utilizing this richer, three-dimensional representation, the predictive architecture acquires a significantly more comprehensive and nuanced understanding of the complex frame-to-frame motion dynamics, including the tracking of intricate object trajectories and the capture of subtle, but critical, structural variations. The ultimate result of this improvement is twofold: on a quantitative level, it directly contributes to an improvement in the accuracy of predictions across various metrics; and on a qualitative, user-perceived level, it improves the perceptual quality of the reconstructed videos, successfully mitigating common visual artifacts such as excessive motion blur, flickering, and the loss of fine texture details, which are often detrimental to the user experience in latency-critical streaming applications.

The Gated Spatiotemporal Attention (GSTA) module functions as the core feature interaction unit. Figure 3 presents a detailed structural comparison between our proposed approach and the baseline. As shown in Figure 3b, the original SimVP-v2 relies on implicit

2D depthwise convolution (DWConv2D). This design processes purely spatial features while treating the temporal dimension passively, forcing the model to infer temporal dependencies through complex, abstract channel-wise operations. Consequently, the resulting features are often disorganized and difficult to interpret, rendering predictions susceptible to motion inconsistency and blurring.

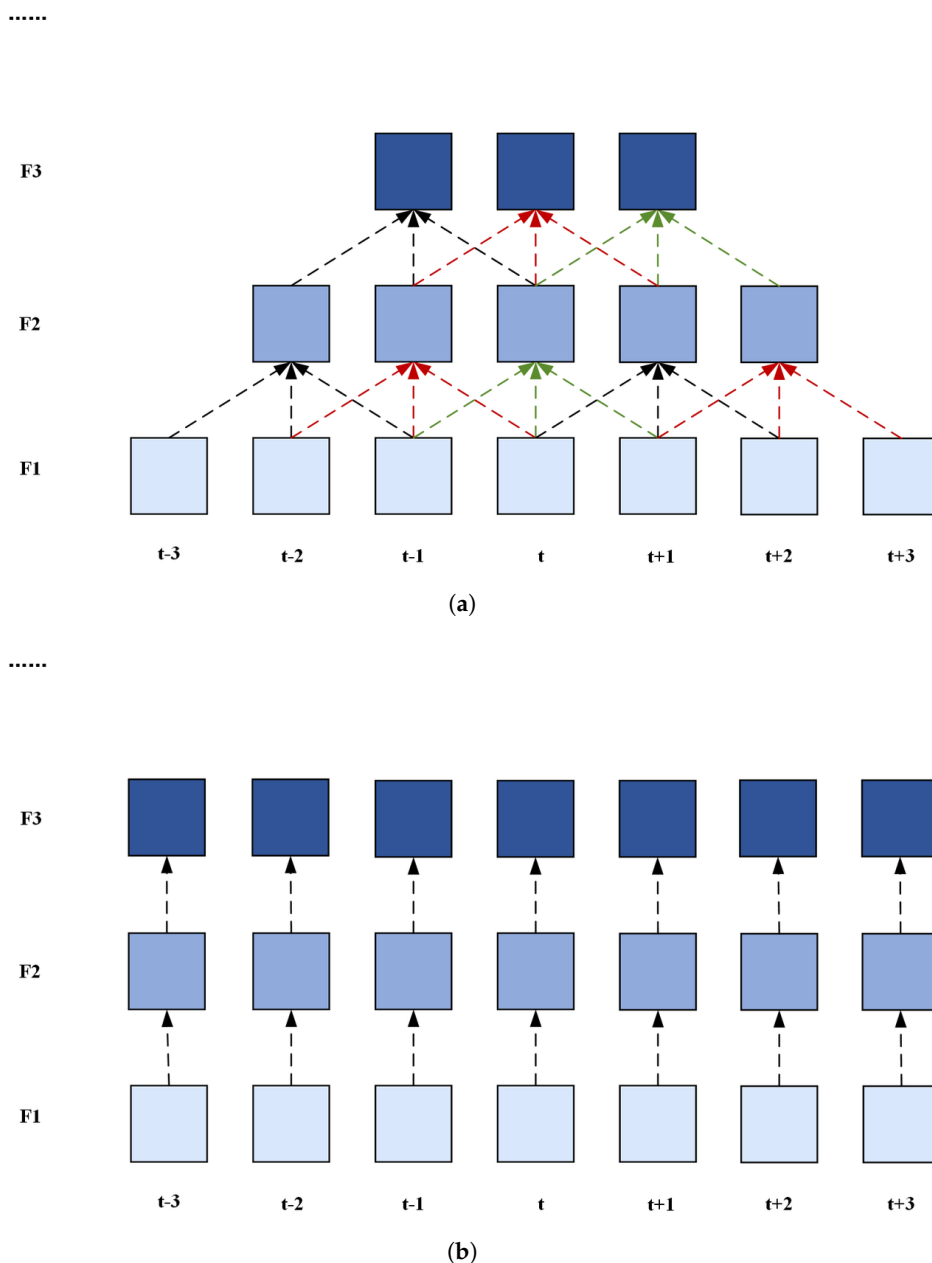


Figure 3. Architectural comparison of the Gated Spatiotemporal Attention (GSTA) modules. (a) Proposed A3DSimVP's GSTA Module; (b) Baseline SimVP-v2's GSTA Module.

In stark contrast, as illustrated in Figure 3a, A3DSimVP integrates explicit 3D depthwise convolution directly within the GSTA block. A defining characteristic of A3DSimVP, which fundamentally distinguishes it from the spatially limited processing of SimVP-v2, is the rapid expansion of its temporal receptive field (RF_T) via this explicit 3D modeling. In this visualization, the increasing network depth is represented by the transition of blue blocks from light to dark shades. Specifically, at the second convolutional stage (the first 3D layer), the model integrates information from three adjacent frames, as indicated by the black, green, and red dashed arrows originating from time steps $t-1$, t , and $t+1$,

respectively. By the third layer, this temporal scope effectively widens to encompass five frames. As the network deepens, the RF_T grows progressively, enabling the aggregation of coherent temporal features across an extended time horizon. This architectural shift ensures that the GSTA module is supplied with explicitly movement-aware representations rather than static spatial maps. Consequently, the model transcends the baseline's limitation of basic frame-to-frame motion recovery, allowing it to robustly capture and model complex, long-range spatiotemporal dependencies.

We define the function RF_T to calculate the maximum number of time steps (frames) our A3DSimVP model can aggregate at the output of the GSTA block's L -th layer of 3D convolution, following a simple stacked architecture. The size of the temporal receptive field at layer L can be calculated using the following formula:

$$RF_T^{(L)} = RF_T^{(L-1)} + (K_T^{(L)} - 1) \times S_T^{(L-1)} \quad (4)$$

For this calculation formula, the definitions of each variable are as follows:

- $RF_T^{(L)}$: Represents the Temporal Receptive Field (the total number of visible time steps) at the output of the L -th 3D convolutional layer.
- $RF_T^{(L-1)}$: The Temporal Receptive Field size from the previous layer ($L - 1$). For the very first layer ($L = 1$), the base receptive field $RF_T^{(0)}$ is 1 (representing a single initial feature map, though the first kernel immediately expands it).
- $K_T^{(L)}$: The temporal dimension of the 3D convolution kernel (filter) at layer L . In our description, this is 3 (e.g., a $3 \times k \times k$ kernel).
- $S_T^{(L-1)}$: The effective temporal stride up to the beginning of the current layer L . This is the cumulative product of the strides of all previous 3D convolutional layers ($L - 1$).

This formula clearly demonstrates how the use of the 3D kernel ($K_T > 1$) in A3DSimVP ensures that the temporal dimension is explicitly and cumulatively expanded with each subsequent layer, a fundamental advantage over the purely 2D approach used in the SimVP-v2 baseline.

Formally, the following equations specify the transformation pipeline within the GSTA-3D block. Each operation is mathematically structured to progressively expand the temporal receptive field, ensuring that the convolutional group K not only extracts local motion cues but also provides a principled foundation for the subsequent attention-based modulation.

$$K(\tilde{Z}^j) = \text{Conv3d}_{1 \times 1 \times 1}(\text{Conv3d}^{\text{Dw-d}}(\text{Conv3d}^{\text{Dw}}(\tilde{Z}^j))) \quad (5)$$

Here, $\tilde{Z}^j$ denotes the input tensor of the j -th GSTA-3D layer. $\text{Conv3d}^{\text{Dw}}$ is a depth-wise separable 3D convolution that efficiently captures spatiotemporal features, while $\text{Conv3d}^{\text{Dw-d}}$ is its dilated version, enlarging the receptive field R_{ST} to capture long-range dependencies. The intermediate output F^j is then split along the channel dimension:

$$\text{Split}(F^j) = G, \bar{F}^j \quad (6)$$

where G serves as the gating coefficients and $\bar{F}^j$ is the modulated feature tensor. The final update is obtained through the attention mechanism:

$$\tilde{Z}^{j+1} = \Psi(G) \odot \bar{F}^j \quad \text{where} \quad \Psi(G) = \sigma(G) \quad (7)$$

Here, $\Psi(G)$ denotes the spatiotemporal attention gate, mapped into $[0, 1]$ via the Sigmoid activation function $\sigma(\cdot)$, and $\odot$ represents element-wise multiplication. This

dynamic gating mechanism adaptively selects and reweights features, thereby significantly enhancing the model's spatiotemporal dynamic perception.

3.3. Multimodal Feature Fusion Module

To improve robustness and enrich predictive information, A3DSimVP incorporates a multi-modal feature fusion module F that integrates audio information A as auxiliary guidance. Specifically, the audio stream is first encoded by the audio encoder E_A , producing a fixed-dimensional context vector a :

$$a = E_A(A), \quad a \in \mathbb{R}^{D_a} \quad (8)$$

This context vector is then transformed by the feature modulation network M_{mod} into a channel-wise weight vector w :

$$w = M_{\text{mod}}(a) \in \mathbb{R}^C \quad (9)$$

During fusion, w is broadcast across the spatiotemporal dimensions of the visual feature Z_V , and additive modulation is applied to generate the multi-modal enhanced feature $\tilde{Z}$:

$$\tilde{Z} = Z_V \odot \text{Expand}(w) + Z_V \quad (10)$$

where $\odot$ denotes element-wise multiplication. This mechanism efficiently injects audio context into visual representations, enhancing cross-modal expressiveness. The fused multi-modal feature $\tilde{Z}$ is subsequently fed into the GSTA-3D transformer T for explicit spatiotemporal evolution [28].

The proposed improvements explicit temporal modeling through 3D depthwise convolutions and multimodal fusion with audio, jointly tackle the two primary weaknesses of SimVP-v2. The 3D kernels explicitly encode motion trajectories, mitigating blurring artifacts and enhancing fidelity in long-horizon predictions. Meanwhile, the audio pathway acts as an auxiliary constraint, particularly useful under degraded network conditions where visual information alone may be insufficient. The combination of these strategies allows our framework to achieve superior robustness and generalization, while maintaining efficiency comparable to SimVP-v2.

4. Experiments

In this section, we conduct extensive experiments on three benchmark datasets: KTH, UCF101, and MotionX, to evaluate the performance of our proposed model. We first introduce the experimental setup and the evaluation metrics used, including MSE, MAE, Structure Similarity Index Measure (SSIM) [44], Peak Signal-to-Noise Ratio (PSNR) [44,45], and Learned Perceptual Image Patch Similarity (LPIPS) [46]. Then, we compare our model with several existing approaches ordered by publication year, including ConvLSTM, MMVP, PhyDNet, PredRNN-v2, and SimVP-v2. Finally, we present ablation studies to demonstrate the effectiveness of our method.

4.1. Data Collection

We employ three distinct datasets (KTH, Motion-X, and UCF-101) to evaluate our method and compare it with existing models. Detailed descriptions of these benchmarks are provided below.

1. KTH Action Dataset: The KTH dataset [47], collected by KTH University, serves as a classical benchmark for human action recognition. A total of 25 subjects performed six fundamental actions (walking, jogging, running, boxing, hand waving, hand clapping) across four distinct experimental scenarios: indoor, outdoor, coat wearing and varying

lighting. The dataset consists of 2391 video pairs in .avi format, characterized by a resolution of 160×120 pixels, a frame rate of 30 FPS, and an average duration of 4.0 s. Due to its minimal background clutter and external interference, KTH is particularly effective for evaluating discriminative motion features in early action recognition studies.

2. Motion-X Dataset: The Motion-X dataset [48] is a large-scale benchmark designed for 3D human motion modeling, distinguished by its diversity of subjects and scenarios. It contains 81,100 video pairs (.avi format) with frame-level annotations, recorded at a resolution of 320×240 pixels and 30 FPS, with an average duration of 6.4 s. In this study, we exclusively utilize the RGB video component, intentionally excluding skeleton and pose annotations, to assess model robustness under multi-subject and multi-scene conditions. Given that Motion-X includes audio data similar to UCF-101, our framework incorporates an audio–visual fusion mechanism. To prepare the data, the audio tracks (typically sampled at 44.1 kHz or 48 kHz) are converted to mono and segmented to align with the duration of each video frame. We then employ Mel-spectrograms to extract acoustic features, generating a 40-dimensional audio vector for each frame. This approach leverages the complementarity of audio and visual modalities to capture complex spatiotemporal dynamics, thereby refining prediction accuracy and stability.

3. UCF-101: The UCF-101 dataset [49] comprises 101 action categories recorded in unconstrained real-world environments. It covers diverse scenarios, including human–object interactions, sports, musical performances, and interpersonal activities. Unlike KTH, UCF-101 introduces significant challenges such as background clutter, illumination variations, and motion blur. Consequently, it serves as a standard benchmark for evaluating generalization. The dataset contains 13,320 video pairs (.avi format) with an average duration of 7.21 s, and the frame rate is 25 FPS. It shares the same native resolution (320×240) as Motion-X. For the accompanying audio, we preserve the original sampling rates of 44.1 kHz or 48.0 kHz. During preprocessing, audio tracks are converted to mono and segmented to align with each video frame. We then employ Mel-spectrogram analysis to extract acoustic features. This process yields a 40-dimensional audio vector for every frame, enabling the model to leverage audio–visual correlations for enhanced representation.

Regarding data preprocessing, we applied different strategies based on the native characteristics of each dataset. For KTH and MotionX, to ensure consistent feature extraction across varying aspect ratios and to optimize computational efficiency, all video frames were resized to a uniform resolution of 120×212 pixels. In contrast, for UCF-101, we retained the original resolution of 320×240 pixels to preserve the fine-grained details necessary for recognizing complex real-world actions. A comprehensive comparison of the detailed features and parameters for all three datasets is presented in Table 2.

Table 2. Comparison of key statistics and characteristics among KTH, Motion-X, and UCF-101 datasets used in this study.

Feature	KTH	Motion-X	UCF-101
Content	Basic Actions	3D Human Motion	101 Action Classes
Frame Rate	30 FPS	30 FPS	25 FPS
Resolution Used	120×212	120×212	320×240
Support Audio	×	✓	✓
Audio Sample Rate	N/A	44.1/48 kHz	44.1/48 kHz
Data Modality	Grayscale	RGB	RGB
Background	Simple	Complex	Complex

4.2. Experimental Setup

In this study, we conducted model evaluations using a total of approximately 17,000 video samples drawn from three datasets: KTH, Motion-X, and UCF-101. For Motion-X and UCF-101, we strictly adhered to an 8:1:1 random split ratio to ensure balanced data distribution. Specifically, on Motion-X, the training set contained 4552 samples, the validation set comprised 557 samples, and the test set included 556 samples. On UCF-101, the training set was expanded to 5476 samples, with 684 in the validation set and 685 in the test set. This meticulous partitioning guaranteed sufficient data for robust parameter learning while maintaining dedicated validation and test sets for hyperparameter optimization and unbiased performance benchmarking.

For the KTH dataset, the experimental design strictly adhered to the normalized, highly reproducible protocols that have long been adopted by the baseline model SimVP-v2 [28] for video prediction tasks. The dataset's original scale and diversity were meticulously partitioned using a rigorous, non-overlapping data splitting strategy to ensure the independence and fairness of the training, validation, and final testing phases. The model was explicitly tasked with predicting the subsequent 20 future frames based on the preceding 10 observed frames. This design skillfully constructs a high-difficulty challenge intended to comprehensively probe the model's intrinsic ability and performance boundaries in capturing high-dynamic temporal dependencies over long time horizons. In this in-depth exploratory study, we specifically focused on the prediction of 20 future frames (i.e., prediction timesteps spanning from $T = 11$ through $T = 30$), utilizing SSIM and PSNR as the primary quality assessment standards for the systematic and rigorous evaluation of the model's robustness and predictive accuracy.

4.3. Quantitative Analysis

We conducted a comprehensive evaluation of the improved SimVP-v2 model on three widely used datasets of UCF-101, Motion-X, and KTH. We using the following performance metrics:

- **MSE and MAE:** These metrics quantify the differences between the model's predicted values and the ground-truth data, thereby serving as indicators of the prediction accuracy.
- **PSNR:** PSNR is an objective measure of the visual quality of reconstructed frames. Higher PSNR values indicate better fidelity, with values below 20 typically reflecting poor quality, values between 20 and 40 representing moderate quality, and values above 40 approximating near-original quality.
- **SSIM:** SSIM evaluates similarities between images in terms of luminance, contrast, and structure based on the human visual system. Unlike PSNR, SSIM provides a perceptually more consistent measure of image quality, where values closer to 1 indicate greater structural similarity.
- **LPIPS:** LPIPS measures perceptual similarity between images by extracting features through pretrained deep neural networks and comparing them in feature space, thereby offering a fine-grained evaluation of perceptual quality.

The evaluation metrics of the improved model on each dataset are presented in Tables 3–5. For the KTH dataset, the evaluation results are shown in Table 3 and Figure 4.

As shown in Table 3, we introduce the Root Mean Square Error (RMSE) metric to provide a stricter assessment of large prediction errors. Our proposed A3DSimVP ($\theta = 4$) consistently outperforms baseline methods across these key metrics. Compared to the strongest competitor, SimVP-v2, our method reduces MSE by 2.66% and MAE by 7.66%. Concurrently, the RMSE decreases from 6.71 to 6.62. This immediate drop in pixel-level error validates that explicit 3D kernels offer superior precision in capturing motion dynamics compared to implicit recurrent updates.

Table 3. Quantitative performance comparison on the KTH dataset, demonstrating that our proposed A3DSimVP ($\theta = 4$) achieves state-of-the-art results against baseline methods.

Dataset	KTH					
Method	MSE ↓	RMSE ↓	MAE ↓	SSIM ↑	PSNR ↑	LPIPS ↓
Convlstm	47.65	6.90	445.55	0.8977	26.99	0.2669
MMVP	93.15	9.65	611.50	0.7313	23.58	0.4102
Phydnet	91.12	9.55	765.65	0.8322	23.41	0.5016
Predrnn-V2	61.51	7.84	532.91	0.7493	25.18	0.3387
SimVP-v2	45.02	6.71	417.75	0.9049	27.04	0.2524
Ours (only 3D, $\theta = 2$)	44.69	6.68	393.37	0.8315	27.27	0.2136
Ours (only 3D, $\theta = 3$)	43.83	6.62	388.20	0.8320	27.34	0.2150
Ours (only 3D, $\theta = 4$)	43.82	6.62	385.73	0.8314	27.35	0.2193

Note: Bold indicates the best performance; ↑ indicates higher is better; ↓ indicates lower is better.

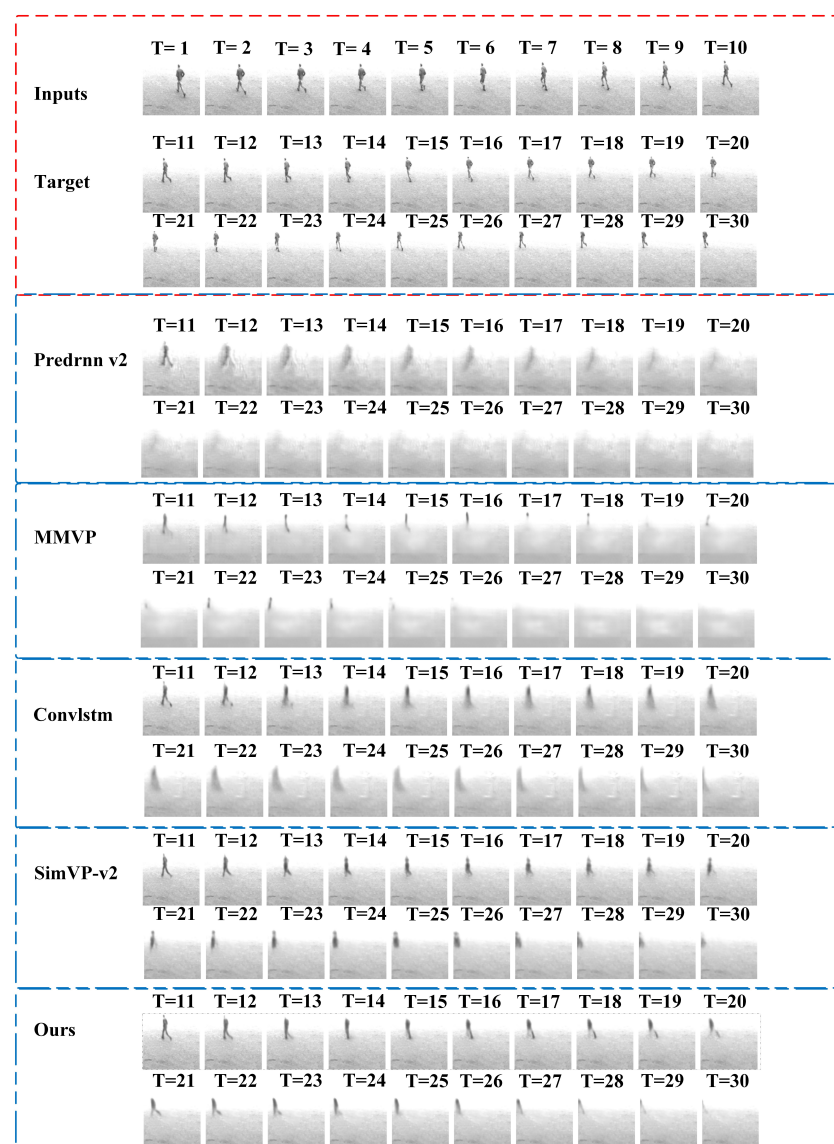


Figure 4. Visual comparison of predicted results on the KTH dataset.

Beyond error minimization, the advantages of our model are evident in perceptual quality. The method achieves a 13.11% reduction in LPIPS compared to SimVP-v2, delivering sharper reconstructions. The combined gains in PSNR (27.35 dB) and LPIPS confirm that

A3DSimVP effectively sustains structural similarity while improving the overall sharpness of the generated frames.

We further investigate the impact of temporal depth by varying θ from 2 to 4. Increasing θ yields consistent improvements in predictive accuracy. Specifically, MSE decreases from 44.69 ($\theta = 2$) to 43.82 ($\theta = 4$), while PSNR rises from 27.27 dB to 27.35 dB. This trend indicates that a larger θ enhances the model's capacity to model complex temporal evolution. Although LPIPS shows minor fluctuations, the configuration with $\theta = 4$ achieves the lowest prediction errors and the highest peak signal-to-noise ratio, offering the optimal balance between pixel fidelity and structural coherence.

The visualization of the KTH dataset is shown in Figure 4.

In Figure 4, T denotes the video time step. The visualization compares the ground truth (Target) against predictions from various baseline models and our proposed A3DSimVP. As illustrated in Figure 4, our method demonstrates higher accuracy by preserving sharp edges and consistent temporal motion, contrasting with the blurrier baseline predictions. This superiority is particularly evident in the later prediction stages (frames $T = 20$ to $T = 23$). While the SimVP-v2 prediction suffers from severe motion blur that renders the subject's legs and body contours almost indistinguishable, A3DSimVP successfully maintains a clear definition of the subject's posture throughout the sequence.

Visual comparisons reinforce these findings, utilizing $T = 1$ to $T = 10$ frames as input and predicting the subsequent $T = 11$ to $T = 30$ frames, where T represents the video time step. Competing approaches, including PredRNN-v2, MMVP, ConvLSTM, and baseline SimVP-v2, frequently suffer from blurriness, detail loss, and poor temporal coherence when applied to dynamic scenes. In contrast, our method captures motion details more precisely and maintains clearer background separation, producing predictions that are both sharper and more temporally consistent.

Quantitatively, our best-performing model, Ours ($\theta = 4$), achieved a state-of-the-art PSNR of 27.352, outperforming the best competing baseline SimVP-v2 (PSNR = 27.041) by approximately 1.15%. Furthermore, focusing on perceptual quality, the Ours ($\theta = 2$) model demonstrated a significant reduction in the LPIPS metric (0.2136) compared to SimVP-v2 (0.2524), representing a 15.37% improvement in perceived image fidelity. By integrating numerical superiority with enhanced perceptual quality, the proposed model demonstrates robust generalization and practicality, offering a strong benchmark for future research in real-time video prediction tasks.

For the MotionX dataset, the evaluation results are shown in Table 4.

Table 4. Quantitative performance comparison on the Motion-X dataset, demonstrating that our proposed A3DSimVP ($\theta = 3$) achieves superior predictive accuracy and perceptual quality compared to baseline methods.

Dataset	Motion-X					
Method	MSE ↓	RMSE ↓	MAE ↓	SSIM ↑	PSNR ↑	LPIPS ↓
Convlstm	423.30	20.57	2637.52	0.8097	24.25	0.1909
MMVP	1221.30	34.95	6190.40	0.3958	18.47	0.7059
Phydnet	887.26	29.79	5396.17	0.5887	20.17	0.4020
Predrnn-V2	420.70	20.51	2568.27	0.8258	24.35	0.1485
SimVP-v2	351.32	18.74	1823.62	0.8846	26.33	0.1161
Ours ($\theta = 3$)	351.07	18.74	1688.96	0.8900	26.52	0.1011

Note: Bold indicates the best performance; ↑ indicates higher is better; ↓ indicates lower is better.

Table 4 details the quantitative evaluation on the Motion-X dataset. To rigorously assess error distribution, we incorporate the RMSE metric. While the MSE (351.07) and RMSE (18.74) are comparable to the competitive SimVP-v2, our method achieves a critical 7.38%

reduction in Mean Absolute Error (MAE), dropping from 1823.62 to 1688.96. This improvement is driven by the explicit 3D spatiotemporal modeling, which enables convolutional kernels to jointly sample across space and time, directly encoding inter-frame displacement and motion trajectories. Consequently, A3DSimVP significantly outperforms recurrent architectures like PredRNN-V2, reducing MAE by 34.24% and offering superior stability.

In terms of perceptual fidelity, the model sets a new standard with a 12.92% reduction in LPIPS (0.1011 vs. 0.1161) and a peak SSIM of 0.8900. This gain is largely attributed to the audio modality fusion, which introduces complementary rhythmic and event-based cues via Channel Attention and lightweight FiLM gating strategies. By enhancing prediction stability at action initiation points, these multimodal signals contribute to the additional reduction in LPIPS. Overall, the Motion-X results provide strong evidence for the ability of our proposed method to simultaneously achieve high prediction accuracy and superior perceptual quality in complex dynamic sequences. Detailed ablation studies analyzing these components are presented in Section 4.4.

For the UCF101 dataset, the evaluation results are shown in Table 5.

Table 5. Quantitative performance comparison on the UCF101 dataset, demonstrating that our proposed A3DSimVP ($\theta = 3$) achieves state-of-the-art performance across all metrics against baseline methods.

Dataset	UCF101					
Method	MSE ↓	RMSE ↓	MAE ↓	SSIM ↑	PSNR ↑	LPIPS ↓
Convlstm	1365.06	36.95	8470.44	0.8236	25.85	0.1690
MMVP	1396.49	37.37	8845.08	0.7185	23.04	0.3126
Phydnet	5410.79	73.56	26,018.02	0.5745	18.25	0.4553
Predrnn-V2	1334.89	36.54	8032.12	0.8355	26.32	0.1482
SimVP-v2	1253.55	35.41	7769.91	0.8414	26.77	0.1417
Ours ($\theta = 2$)	1234.87	35.14	7557.05	0.8348	26.91	0.1340
Ours ($\theta = 3$)	1132.53	33.65	6844.49	0.8453	26.91	0.1324

Note: Bold indicates the best performance; ↑ indicates higher is better; ↓ indicates lower is better.

As detailed in Table 5, the A3DSimVP model demonstrates a decisive performance advantage on the UCF101 dataset. To strictly assess error distribution, we introduce the RMSE metric. Quantitatively, our proposed architecture achieves a substantial reduction in prediction error compared to the SimVP-v2 baseline. Specifically, the MSE is reduced by 9.65% (from 1253.55 to 1132.53), and the RMSE drops from 35.41 to 33.65. Concurrently, the MAE decreases by 11.91% (from 7769.91 to 6844.49). These significant drops in error metrics indicate that the integration of explicit 3D spatiotemporal modeling effectively mitigates the cumulative deviations often observed in rapid motion sequences. By capturing long-range temporal dependencies more accurately than the 2D-based baseline, our model ensures superior pixel-level alignment and stability across extended prediction horizons.

Beyond minimizing pixel-level errors, A3DSimVP excels in maintaining structural integrity and perceptual fidelity. The model attains a PSNR of 26.91 dB and an SSIM of 0.8453. Most critically, the Learned Perceptual LPIPS score is reduced to 0.1324. This represents a 6.6% improvement over the SimVP-v2 baseline (0.1417). The superior LPIPS performance suggests that the auxiliary audio cues effectively guide the generation process, reducing blurring artifacts and resulting in perceptually sharper, more realistic video content even in complex, dynamic scenes.

The results from the qualitative assessment powerfully corroborate the exceptional performance demonstrated by the quantitative metrics of our method. As illustrated in Figure 5, we provide a detailed visual comparison of video prediction outputs from our model alongside various competing models on the UCF-101 dataset. The prediction

sequences shown progress across time steps $T = 1, 2, 3, 4$, and 5 , where T denotes the index of the predicted future frame. The successful maintenance of detail and motion coherence, particularly at the later time step, $T = 5$, serves as a crucial, high-stakes test for the model's ability to maintain predictive quality over challenging long-term motion dependencies and resist the rapid accumulation of prediction error. Our visual outputs confirm that the explicit 3D modeling translates into clearer boundaries, fewer ghosting artifacts, and superior preservation of fine texture details over extended temporal horizons. Detailed ablation experiments regarding this dataset are discussed in Section 4.4.

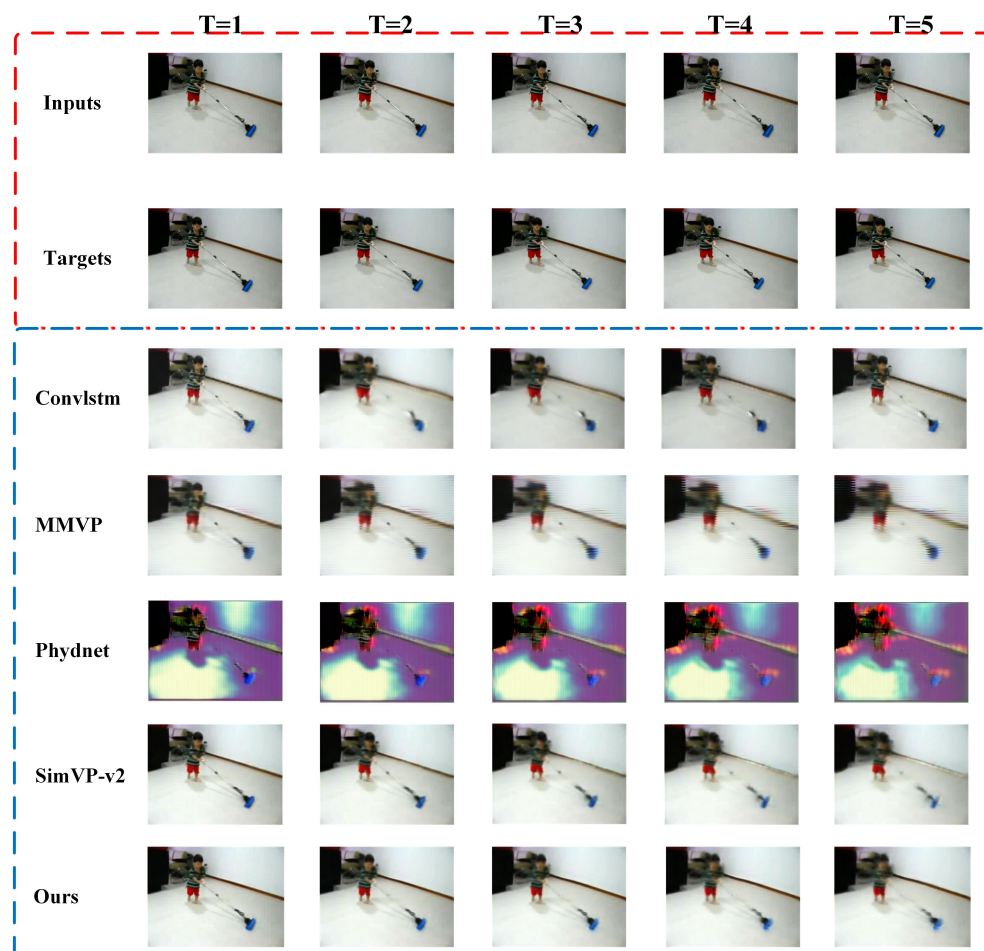


Figure 5. Visual comparison of predicted results on the UCF101 dataset. The models predict subsequent frames (Targets) based on the first $T = 5$ input frames. The red dashed rectangle indicates the ground truth frames, and the blue dashed rectangle highlights the prediction results generated by different methods.

As illustrated in Figure 5, we present a qualitative comparison between A3DSimVP and baseline models, including ConvLSTM and MMVP. The visualization demonstrates that our method achieves superior visual fidelity across the prediction horizon. Specifically, while competing models frequently suffer from significant motion blur and loss of fine texture, A3DSimVP effectively minimizes blurring artifacts and preserves sharp structural details. This qualitative evidence corroborates the quantitative results, highlighting the model's capability to maintain clarity in complex motion scenarios.

In the prediction example of the Base Model, we observe that most comparative methods exhibit clear shortcomings in prediction quality. For instance, the predictions generated by the MMVP and ConvLSTM models are generally plagued by severe defocus blurring and loss of high-frequency details. The contours of moving objects, such as the hands and

tools, become noticeably blurred after $T = 5$. The PhyDNet model often introduces pronounced artifacts and structural distortions. Even our Base Model, SimVP-V2, an efficient all-convolutional method designed to decouple spatial and temporal dependencies, shows a clear degradation in image quality during long-term prediction ($T = 3$ to $T = 5$), struggling to sustain stable sharpness.

Our method has superior visual effects compared to other methods in terms of visual perception, and our prediction results have better visual effects:

- **Exceptional Detail Preservation:** Our method precisely retains fine textures and sharp edges that other models fail to maintain. At the most challenging time step, $T = 9$, the prediction images from other models are often too blurred to discern the geometric shape of the object or subtle background textures, such as leg contours or the edges of the tool. Our model, however, consistently provides a clear rendering of the tool's geometric structure, minute floor texture details, and crisp object boundaries.
- **Superior Motion Coherence:** Our predicted frames exhibit exceptional temporal coherence, effectively eliminating the flickering and structural inconsistencies commonly observed in competitor sequences, thereby producing a visually smoother and more realistic video flow.
- **Prediction Stability:** As the time step T increases, the drop in clarity for our model's predictions is significantly less severe than that of any baseline model, confirming the powerful capability of our approach in suppressing error accumulation and maintaining stability during long-range forecasting.

In summary, the performance of our method in this qualitative assessment, characterized by superior sharpness, more accurate texture details, and stable long-term prediction, is in perfect agreement with the quantitative improvements, definitively confirming the efficacy and superiority of our introduced modules and strategies in complex video prediction tasks.

4.4. Ablation Studies and Component Analysis

To systematically quantify the distinct contributions of the proposed 3D temporal modeling module and the audio modality fusion strategy, we conducted comprehensive ablation studies. The experimental design evaluates three critical configurations: the standard SimVP-v2 baseline, a variant integrating only the 3D depthwise convolution (Only 3D), and the complete dual-modality model (Ours).

We first present the comparative results on the UCF101 dataset in Table 6. The introduction of the 3D module alone yields a reduction in MSE from 1253.55 to 1248.19 compared to the baseline. This result validates the role of explicit 3D temporal modeling in enhancing inter-frame dependency capture. Furthermore, the full model ('Ours') achieves the most robust performance with the lowest MSE of 1243.87 and an optimal LPIPS score of 0.1340. The improvement in LPIPS signifies that the audio signal acts as a strong auxiliary cue, boosting prediction accuracy and enhancing perceptual realism.

Table 6. Ablation study on the UCF101 dataset quantifying the individual and combined contributions of 3D convolution and audio features.

Method	MSE ↓	MAE ↓	SSIM ↑	PSNR ↑	LPIPS ↓
SimVP-v2 (Base)	1253.55	7769.91	0.8414	26.77	0.1417
Only 3D ($\theta = 3$)	1248.19	7661.05	0.8435	26.86	0.1399
Only Audio ($\theta = 3$)	1250.95	7733.80	0.8415	26.76	0.1446
Ours	1243.87	7557.05	0.8348	26.91	0.1340

Note: Bold indicates the best performance; ↑ indicates higher is better; ↓ indicates lower is better.

Regarding the KTH dataset, it is important to note that it lacks audio modalities. Consequently, experiments on KTH were strictly limited to verifying the structural enhancements of the 3D module (as detailed in the main results). This distinction explains why the multimodal ablation analysis presented here for UCF101 and Motion-X cannot be replicated for KTH.

To further validate our architecture on a large-scale human motion domain, we extended our ablation analysis to the Motion-X dataset. As shown in Table 7, the ‘Only 3D’ variant demonstrates a strong capability in enhancing structural fidelity, significantly reducing MAE by 6.09% (from 1823.62 to 1712.60) and LPIPS by 12.75% (from 0.1161 to 0.1013) compared to the baseline. Although this variant exhibits a slight fluctuation in MSE (359.38 versus 351.32), the higher SSIM (0.8885) indicates that explicit 3D modeling prioritizes structural coherence. The ‘Only Audio’ variant provides marginal gains when used in isolation. However, the Full Model (‘Ours’) effectively integrates both strengths. It corrects the MSE regression observed in the 3D-only model, achieving the lowest MSE of 351.07, while simultaneously attaining the best MAE (1688.96) and LPIPS (0.1011) scores. This confirms that the audio context serves as a stabilizing constraint, complementing the 3D spatial-temporal features to generate predictions that are both accurate and perceptually realistic.

Table 7. Ablation study on the Motion-X dataset quantifying the impact of 3D convolution and audio fusion.

Method	MSE ↓	MAE ↓	SSIM ↑	PSNR ↑	LPIPS ↓
SimVP-v2 (Base)	351.32	1823.62	0.8846	26.33	0.1161
Only 3D ($\theta = 3$)	359.38	1712.60	0.8885	26.39	0.1013
Only Audio ($\theta = 3$)	351.34	1834.10	0.8847	26.40	0.1115
Ours	351.07	1688.96	0.8900	26.52	0.1011

Note: Bold indicates the best performance; ↑ indicates higher is better; ↓ indicates lower is better.

Having established the efficacy of the components, we proceeded to determine the optimal method for integrating these audio–visual modalities. We evaluated three distinct fusion mechanisms on the Motion-X dataset: simple additive fusion (concatenation), Squeeze-and-Excitation (SE) fusion, and our proposed FiLM-based fusion. As shown in Table 8, the simple Additive method yields the highest error rates (MSE: 631.64), suggesting that direct superimposition fails to capture complex non-linear correlations. The SE-Fusion strategy shows marginal improvement but lags in perceptual quality. In stark contrast, our FiLM strategy achieves a decisive performance leap, reducing the MSE to 351.07 and the MAE to 1688.96, while attaining the lowest LPIPS (0.1011). These results confirm that the affine transformation mechanism in FiLM allows the audio signal to dynamically modulate visual features, effectively suppressing ambiguity.

Table 8. Comparison of different audio–visual fusion strategies on the Motion-X dataset. All models are trained with $\theta = 3$ and an input resolution of 120×212 . The best results are highlighted in bold.

Method	MSE ↓	MAE ↓	SSIM ↑	PSNR ↑	LPIPS ↓
Additive	631.64	2890.14	0.8207	23.10	0.1349
SE Fusion	629.89	2902.66	0.8207	23.10	0.1358
FiLM (Ours)	351.07	1688.96	0.8900	26.52	0.1011

Note: Bold indicates the best performance; ↑ indicates higher is better; ↓ indicates lower is better.

Finally, acknowledging that efficiency is a core value of the SimVP architecture, we analyzed the computational cost. Although the introduction of 3D convolutions and the FiLM audio module increases the parameter count, the impact on inference latency is

effectively managed through a lightweight design. Specifically, our model maintains a competitive inference speed of 20.86 FPS with a computational load of 138 GFLOPs. This strikes an optimal balance between high-fidelity prediction and resource efficiency, ensuring that the significant improvements in perceptual quality do not come at a prohibitive computational expense.

4.5. Computing Resource Consumption Analysis

To evaluate the deployment feasibility of A3DSimVP, we benchmarked model complexity and inference speed against the baseline SimVP-v2 and representative recurrent architectures (ConvLSTM, PredRNN-v2, and MMVP). While we conducted efficiency benchmarks across all datasets, we present the representative analysis on the large-scale Motion-X dataset to demonstrate the deployment feasibility of A3DSimVP. The task involved predicting 20 future frames based on 10 input frames. Inference throughput was measured on a single NVIDIA GPU using a batch size of 1, with the input resolution standardized to 120×212 .

Table 9 presents the detailed resource consumption metrics. Compared to the SimVP-v2 baseline, A3DSimVP incurs a 15.0% increase in FLOPs (0.120T to 0.138T) and requires 44.8% more GPU memory (10,244 MiB to 14,830 MiB). This moderate increase in computational load is primarily due to the introduction of explicit 3D convolutions for spatiotemporal modeling.

Table 9. Comparison of model complexity and inference efficiency on the Motion-X dataset (120×212). ‘Efficiency’ denotes inference throughput (samples/second).

Model	Params (M)	FLOPs (T)	Efficiency (Sample/s) ↑	GPU Mem (MiB)
ConvLSTM	15.497	0.557	8.35	2942
MMVP	0.467	0.169	2.28	38,956
PhyDNet	3.093	0.145	4.51	2338
PredRNN-v2	24.576	1.138	3.16	4808
SimVP-v2	15.634	0.120	12.75	10,244
Ours ($\theta = 3$)	15.667	0.138	20.86	14,830

Note: Bold indicates the best performance; ↑ indicates higher is better.

However, this resource investment yields a substantial return in processing speed. The inference efficiency rises by 63.6%, increasing from 12.75 samples/s to 20.86 samples/s. This result indicates that our architecture effectively utilizes the parallel computing capabilities of GPUs, avoiding the serial bottlenecks often found in recurrent units or unoptimized 2D implementations. Furthermore, compared to the complex recurrent model PredRNN-v2, our method reduces FLOPs by approximately 88% while delivering a $6.6\times$ speedup.

In summary, the quantitative results confirm that A3DSimVP presents a highly favorable trade-off between resource consumption and computational efficiency. By accepting a manageable increase in GPU memory and FLOPs relative to SimVP-v2, our method achieves significantly higher inference throughput. This balance supports the suitability of A3DSimVP for real-time video prediction tasks where low latency is critical.

5. Discussion

Although the proposed A3DSimVP framework demonstrates significant performance improvements in spatiotemporal prediction by enhancing SimVP-v2 through the use of 3D convolutions, its design still exhibits inherent limitations. The first major issue lies in the trade-off between explicit temporal modeling and computational efficiency. Specifically, the replacement of 2D convolutions with 3D convolutions greatly increases computational complexity and memory requirements. As a result, the current architecture can stably

support input resolutions only up to 320×320 , which restricts its applicability in high-definition or ultra-high-definition video prediction scenarios.

The second limitation stems from the multimodal fusion mechanism. While the introduction of audio information improves robustness, the current lightweight, non-adaptive early-fusion design relies mainly on channel modulation. This design lacks the ability to explicitly align audio–visual events in the temporal dimension. Consequently, when minor temporal misalignments occur between modalities, or when the relevance of audio to visual prediction is low, the model cannot dynamically adjust modality weighting, leading to the potential introduction of noise or irrelevant cues.

A further limitation is that the convolutional modules repeatedly extract redundant information from similar frames, which reduces feature diversity. One possible direction is incorporating positional encoding into convolution operations, enabling the model to capture richer spatial–temporal cues. This approach is feasible, as it has been effectively applied in vision transformers with low additional cost.

6. Conclusions

In this study, we present A3DSimVP (Audio- and 3D-Enhanced SimVP-v2), a novel video prediction framework [17] that systematically advances the state of the art in spatiotemporal modeling for real-time and high-fidelity applications. The design of A3DSimVP is motivated by two key challenges in existing methods: the reliance on 2D convolutions for implicit temporal aggregation, which often leads to motion blur and detail loss, and the single-modality constraint that ignores the complementary contextual cues embedded in audio streams. To overcome these limitations, we introduce two fundamental innovations that significantly extend the capabilities of the SimVP-v2 baseline while retaining its hallmark efficiency.

First, we enhance temporal modeling by replacing all 2D operators with 3D depthwise separable convolutions and introducing a redesigned Gated Spatiotemporal Attention (GSTA-3D) module. This explicit spatiotemporal representation enables the model to directly capture coherent motion trajectories and long-term temporal dependencies, substantially improving its ability to preserve fine-grained texture and edge sharpness in predicted frames. Unlike implicit 2D temporal modeling, our 3D approach delivers robust improvements in both predictive accuracy and perceptual clarity, ensuring structural fidelity even in scenarios involving complex motion or long prediction horizons.

Second, we incorporate an efficient audio–visual fusion strategy that leverages synchronized audio features as an auxiliary modality. Through lightweight channel attention and FiLM-based gating, A3DSimVP integrates audio cues with visual representations in a dynamic and context-aware manner. This multimodal enhancement enriches the model’s predictive capacity by providing additional semantic and temporal context, particularly valuable when visual information alone is ambiguous or insufficient. The fusion mechanism strengthens environmental awareness, enhances scene understanding, and contributes directly to the improved perceptual realism of generated frames.

Extensive experimental validation across multiple benchmark datasets demonstrates the superiority of A3DSimVP over both SimVP-v2 and other state-of-the-art approaches. Quantitative results reveal consistent reductions in MSE and MAE, confirming significant decreases in pixel-level prediction errors, while higher SSIM and PSNR scores substantiate the model’s ability to maintain structural similarity and visual fidelity. Most notably, the substantial reduction in LPIPS highlights the enhanced perceptual quality and realism of predicted sequences. Together, these results establish A3DSimVP as a robust, efficient, and high-performing solution capable of meeting the stringent demands of modern latency-

critical applications such as live streaming, immersive communications, video conferencing, and beyond.

Author Contributions: Conceptualization, investigation, resources, funding acquisition, J.Y.; investigation, writing—original draft preparation, writing—review and editing, project administration, H.Z.; resources, funding acquisition, L.L.; writing—review and editing, W.C.; data curation, Q.L.; visualization, H.P.; software, validation, formal analysis, writing, M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key R&D Program of China under Grant 2023YFC3305002, NSFC under Grant 62376092, Natural Science Foundation of Hunan Province under Grant 2024JJ5113, in part by the Research Foundation of Education Bureau of Hunan Province under Grant 23A0462 and 23B0592.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Khan, M.A.; Baccour, E.; Chkirbene, Z.; Erbad, A.; Hamila, R.; Hamdi, M.; Gabbouj, M. A survey on mobile edge computing for video streaming: Opportunities and challenges. *IEEE Access* **2022**, *10*, 120514–120550. [\[CrossRef\]](#)
2. Liu, J.; Mi, Y.; Zhang, X.; Li, X. Task graph offloading via deep reinforcement learning in mobile edge computing. *Future Gener. Comput. Syst.* **2024**, *158*, 545–555. [\[CrossRef\]](#)
3. Frossard, P.; Verscheure, O. Joint source/FEC rate selection for quality-optimal MPEG-2 video delivery. *IEEE Trans. Image Process.* **2001**, *10*, 1815–1825. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Yu, C.; Xu, Y.; Liu, B.; Liu, Y. “Can you SEE me now?” A measurement study of mobile video calls. In Proceedings of the IEEE INFOCOM 2014—IEEE Conference on Computer Communications, Toronto, ON, Canada, 27 April–2 May 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 1456–1464.
5. Afzal, S.; Testoni, V.; Rothenberg, C.E.; Kolan, P.; Bouazizi, I. A holistic survey of multipath wireless video streaming. *J. Netw. Comput. Appl.* **2023**, *212*, 103581. [\[CrossRef\]](#)
6. Wang, Y.; Zhu, Q.F. Error control and concealment for video communication: A review. *Proc. IEEE* **1998**, *86*, 974–997. [\[CrossRef\]](#)
7. Sankisa, A.; Punjabi, A.; Katsaggelos, A.K. Video error concealment using deep neural networks. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 380–384.
8. Höppe, T.; Mehrjou, A.; Bauer, S.; Nielsen, D.; Dittadi, A. Diffusion models for video prediction and infilling. *arXiv* **2022**, arXiv:2206.07696. [\[CrossRef\]](#)
9. Ulhaq, A.; Akhtar, N. Efficient diffusion models for vision: A survey. *arXiv* **2022**, arXiv:2210.09292.
10. Zhao, W.; Rao, Y.; Liu, Z.; Liu, B.; Zhou, J.; Lu, J. Unleashing text-to-image diffusion models for visual perception. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 5729–5739.
11. Deng, Z.; He, X.; Peng, Y.; Zhu, X.; Cheng, L. MV-Diffusion: Motion-aware video diffusion model. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 7255–7263.
12. Wang, Z.; Li, D.; Wu, Y.; He, T.; Bian, J.; Jiang, R. Diffusion models in 3d vision: A survey. *arXiv* **2024**, arXiv:2410.04738. [\[CrossRef\]](#)
13. Croitoru, F.A.; Hondru, V.; Ionescu, R.T.; Shah, M. Diffusion models in vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10850–10869. [\[CrossRef\]](#)
14. Teed, Z.; Deng, J. Raft: Recurrent all-pairs field transforms for optical flow. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 402–419.
15. Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; Brox, T. FlowNet: Learning optical flow with convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2758–2766.
16. Shi, X.; Huang, Z.; Bian, W.; Li, D.; Zhang, M.; Cheung, K.C.; See, S.; Qin, H.; Dai, J.; Li, H. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 12469–12480.
17. Wei, H.; Yin, X.; Lin, P. Novel video prediction for large-scale scene using optical flow. *arXiv* **2018**, arXiv:1805.12243. [\[CrossRef\]](#)

18. Chen, Y.; Zhu, X.; Li, T. A physical coherence benchmark for evaluating video generation models via optical flow-guided frame prediction. *arXiv* **2025**, arXiv:2502.05503. [[CrossRef](#)]
19. Brox, T.; Bruhn, A.; Papenberger, N.; Weickert, J. High accuracy optical flow estimation based on a theory for warping. In Proceedings of the European Conference on Computer Vision, Prague, Czech Republic, 11–14 May 2004; Springer: Berlin/Heidelberg, Germany, 2004; pp. 25–36.
20. Sun, D.; Roth, S.; Black, M.J. Secrets of optical flow estimation and their principles. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 2432–2439.
21. Sun, D.; Yang, X.; Liu, M.Y.; Kautz, J. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8934–8943.
22. Dong, J.; Ota, K.; Dong, M. Video frame interpolation: A comprehensive survey. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 1–31. [[CrossRef](#)]
23. Bao, W.; Lai, W.S.; Zhang, X.; Gao, Z.; Yang, M.H. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 933–948. [[CrossRef](#)] [[PubMed](#)]
24. Meyer, S.; Djelouah, A.; McWilliams, B.; Sorkine-Hornung, A.; Gross, M.; Schroers, C. Phasenet for video frame interpolation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 498–507.
25. Parihar, A.S.; Varshney, D.; Pandya, K.; Aggarwal, A. A comprehensive survey on video frame interpolation techniques. *Vis. Comput.* **2022**, *38*, 295–319. [[CrossRef](#)]
26. Tulyakov, S.; Gehrig, D.; Georgoulis, S.; Erbach, J.; Gehrig, M.; Li, Y.; Scaramuzza, D. Time lens: Event-based video frame interpolation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 16155–16164.
27. Sim, H.; Oh, J.; Kim, M. Xvfi: Extreme video frame interpolation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 14489–14498.
28. Tan, C.; Gao, Z.; Li, S.; Li, S.Z. SimVPv2: Towards simple yet powerful spatiotemporal predictive learning. *IEEE Trans. Multimed.* **2025**, *27*, 5170–5184. [[CrossRef](#)]
29. Wang, Y.; Wu, H.; Zhang, J.; Gao, Z.; Wang, J.; Yu, P.S.; Long, M. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2208–2225. [[CrossRef](#)]
30. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241. [[CrossRef](#)]
31. Huo, Y.; Gang, S.; Guan, C. FCIHMRT: Feature cross-layer interaction hybrid method based on Res2Net and transformer for remote sensing scene classification. *Electronics* **2023**, *12*, 4362. [[CrossRef](#)]
32. Suin, M.; Rajagopalan, A. Gated spatio-temporal attention-guided video deblurring. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 7802–7811.
33. Chicco, D.; Warrens, M.J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [[CrossRef](#)]
34. Wah, B.W.; Su, X.; Lin, D. A survey of error-concealment schemes for real-time audio and video transmissions over the Internet. In Proceedings of the International Symposium on Multimedia Software Engineering, Taipei, Taiwan, 11–13 December 2000; IEEE: Piscataway, NJ, USA, 2000; pp. 17–24.
35. Benice, R.; Frey, A. An analysis of retransmission systems. *IEEE Trans. Commun. Technol.* **2003**, *12*, 135–145. [[CrossRef](#)]
36. Tan, C.; Li, S.; Gao, Z.; Guan, W.; Wang, Z.; Liu, Z.; Wu, L.; Li, S.Z. Openstl: A comprehensive benchmark of spatio-temporal predictive learning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 69819–69831.
37. Oprea, S.; Martinez-Gonzalez, P.; Garcia-Garcia, A.; Castro-Vargas, J.A.; Orts-Escolano, S.; Garcia-Rodriguez, J.; Argyros, A. A review on deep learning techniques for video prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2806–2826. [[CrossRef](#)]
38. Wen, C.; Liu, S.; Yao, X.; Peng, L.; Li, X.; Hu, Y.; Chi, T. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. *Sci. Total Environ.* **2019**, *654*, 1091–1099. [[CrossRef](#)]
39. Gao, Z.; Tan, C.; Wu, L.; Li, S.Z. Simvp: Simpler yet better video prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3170–3180.
40. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 12179–12188.
41. Tan, C.; Gao, Z.; Wu, L.; Xu, Y.; Xia, J.; Li, S.; Li, S.Z. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18770–18782.

42. Lu, J.; Clark, C.; Lee, S.; Zhang, Z.; Khosla, S.; Marten, R.; Hoiem, D.; Kembhavi, A. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 26439–26455.
43. Wei, Y.; Hu, D.; Tian, Y.; Li, X. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv* **2022**, arXiv:2208.09579. [[CrossRef](#)]
44. Hore, A.; Ziou, D. Image quality metrics: PSNR vs. SSIM. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 3–26 August 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 2366–2369.
45. Li, L.; Song, S.; Lv, M.; Jia, Z.; Ma, H. Multi-Focus Image Fusion Based on Fractal Dimension and Parameter Adaptive Unit-Linking Dual-Channel PCNN in Curvelet Transform Domain. *Fractal Fract.* **2025**, *9*, 157. [[CrossRef](#)]
46. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
47. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition—ICPR 2004, Cambridge, UK, 23–26 August 2004; IEEE: Piscataway, NJ, USA, 2004; Volume 3, pp. 32–36.
48. Lin, J.; Zeng, A.; Lu, S.; Cai, Y.; Zhang, R.; Wang, H.; Zhang, L. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 25268–25280.
49. Soomro, K.; Zamir, A.R.; Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.