CVPR Nashville JUNE 11-15, 2025

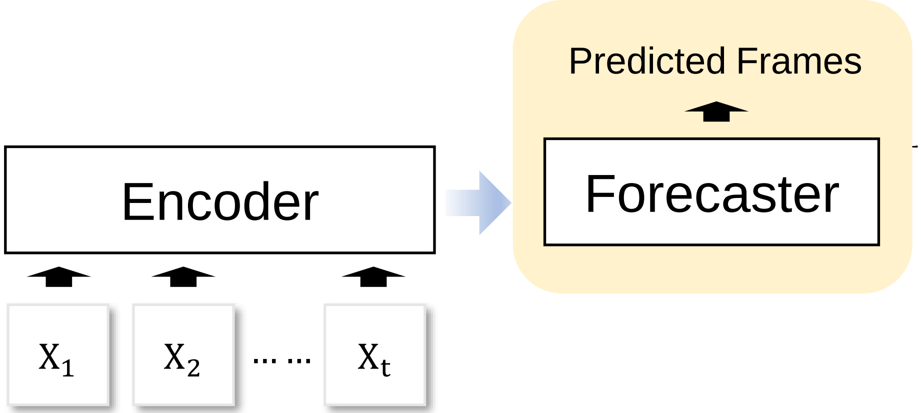
# SRVP:
# Strong Recollection Video Prediction Model Using Attention-Based Spatiotemporal Correlation Fusion

Yuseon Kim Kyongseok Park

KISTI
www.kisti.re.kr
Korea Institute of
Science and Technology Information

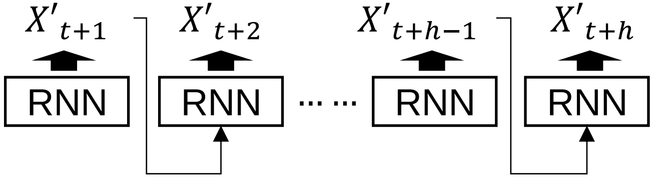
UST
UNIVERSITY OF
SCIENCE & TECHNOLOGY

# Video Prediction



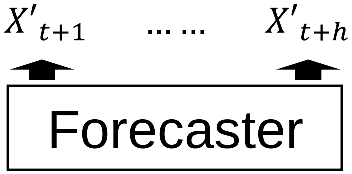
Historical Frames

## RNN-Based Approach



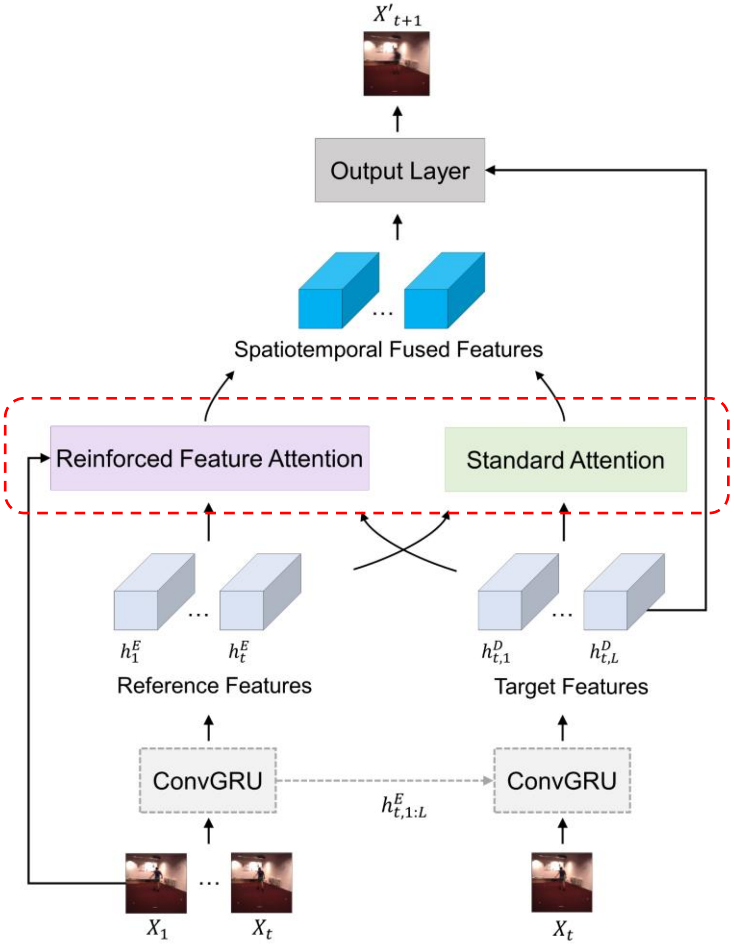
- ✓ Effectively Capture Temporal Changes
- ✓ ConvLSTM based Models
- ✓ Highly Sensitive to Object Motion
- ✓ Blurred Images

## RNN-Free Approach



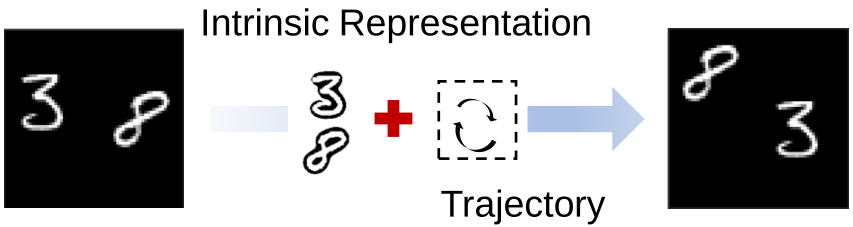
- ✓ Mitigates Cumulative Errors
- ✓ Vision Transformer based Models
- ✓ Absence of Explicit Mechanism for Temporal Variation

# Strong Recollection Video Prediction



- ✓ Baseline Architecture
  - ConvGRU based Encoder-Forecaster
- ✓ Temporal Dynamics

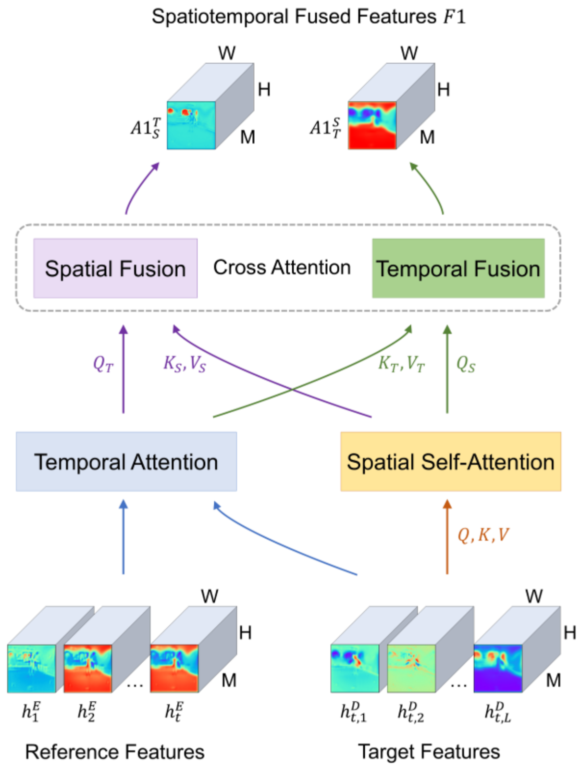
- ✓ Spatiotemporal Fused Features



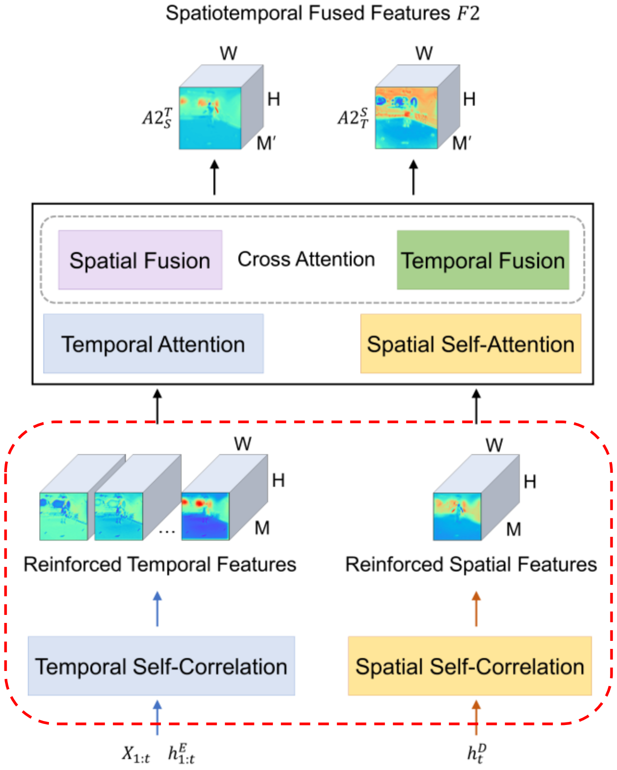
- ✓ Scaled-Dot Product Attention
  - Vaswani, Ashish, et al. "Attention is all you need." In NeurIPS (2017).

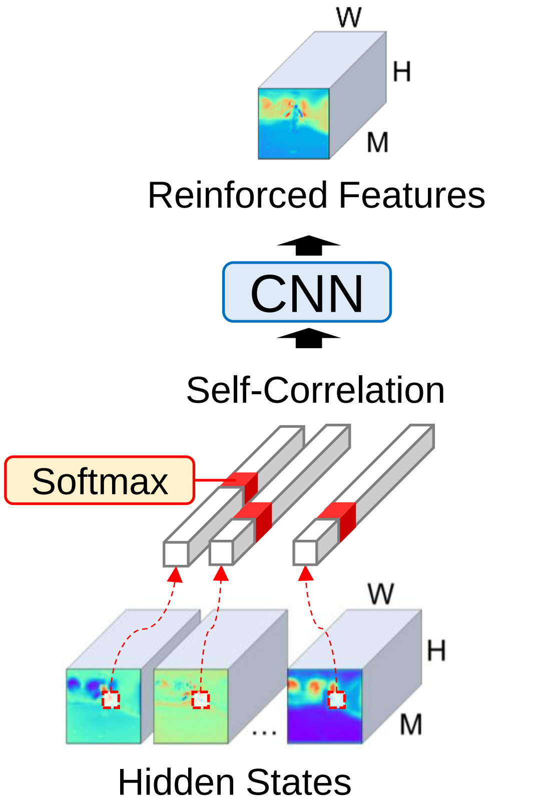
# Attention-based Spatiotemporal Fusion

## Standard Attention



## Reinforced Feature Attention





# Experiment Materials

## ✓ Benchmarks

**Moving MNIST**
- 64x64 Gray Image
- 10 Frames → 10 Frames

**KTH Action**
- 64x64 Gray Image
- 10 Frames → 10 Frames

**Human3.6M**
- 100x100 Color Image
- 4 Frames → 4 Frames

## ✓ Evaluation Metrics

- Mean Squared Error (MSE) ↓
- Peak Signal-to-Noise Ratio (PSNR) ↑
- Structural Similarity Index Measure (SSIM) ↑

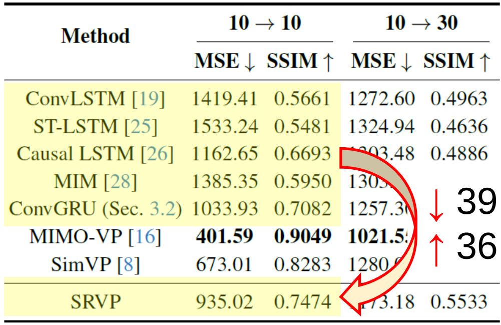
## ✓ Reference Models

| Approach | Model | Conference |
|---|---|---|
| RNN-based | ConvLSTM | NIPS 2015 |
| | ST-LSTM | NIPS 2017 |
| | Causal LSTM | ICML 2018 |
| | MIM | CVPR 2019 |
| | ConvGRU | - |
| RNN-free | MIMO-VP | AAAI 2023 |
| | SimVP | CVPR 2022 |

- MIMO-VP → Transformer-based VP
- SimVP → CNN-based VP

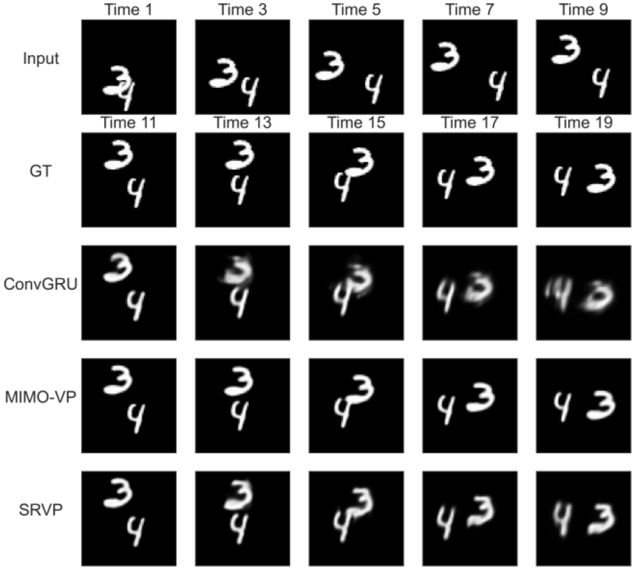
# Notable Findings on SRVP

## ✓ Outperform than RNN-based Models



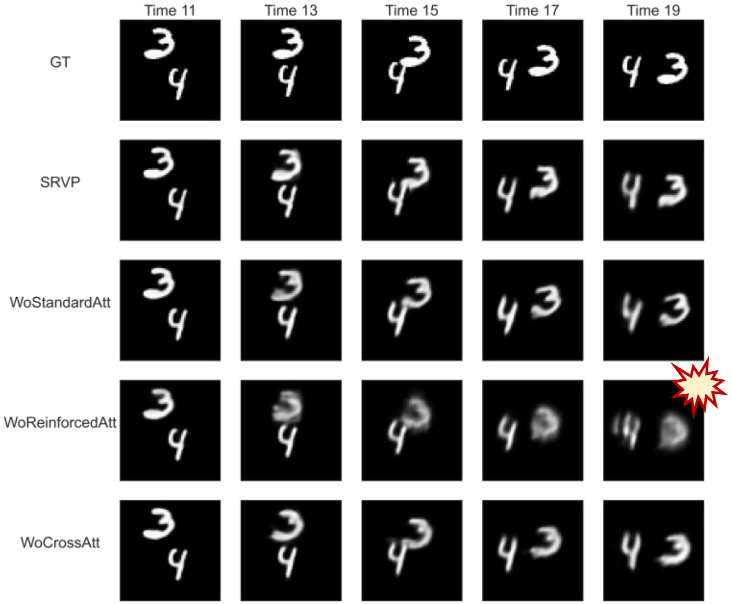
| Method | 10 → 10 | | 10 → 30 | |
|---|---|---|---|---|
| | MSE ↓ | SSIM ↑ | MSE ↓ | SSIM ↑ |
| ConvLSTM [19] | 1419.41 | 0.5661 | 1272.60 | 0.4963 |
| ST-LSTM [25] | 1533.24 | 0.5481 | 1324.94 | 0.4636 |
| Causal LSTM [26] | 1162.65 | 0.6693 | [illegible]03.48 | 0.4886 |
| MIM [28] | 1385.35 | 0.5950 | 1303.[illegible] | |
| ConvGRU (Sec. 3.2) | 1033.93 | 0.7082 | 1257.3[illegible] | |
| MIMO-VP [16] | **401.59** | **0.9049** | **1021.5[illegible]** | |
| SimVP [8] | 673.01 | 0.8283 | 1280.[illegible] | |
| SRVP | 935.02 | 0.7474 | [illegible]73.18 | 0.5533 |

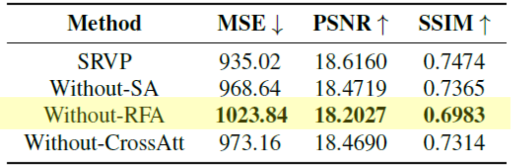
↓ 39.02% in MSE
↑ 36.36% in SSIM



↑ 6.3% in FLOPs
↓ 9.6% in MSE

## ✓ Impact of Feature Reinforcement

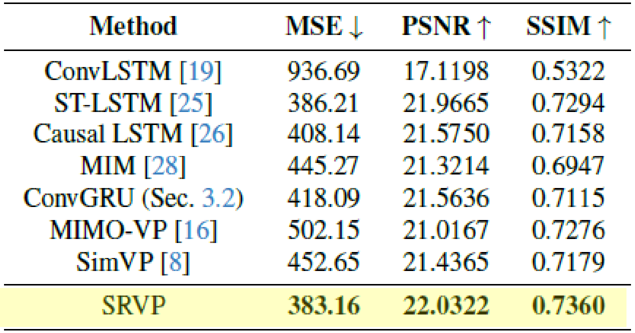




| Method | MSE ↓ | PSNR ↑ | SSIM ↑ |
|---|---|---|---|
| SRVP | 935.02 | 18.6160 | 0.7474 |
| Without-SA | 968.64 | 18.4719 | 0.7365 |
| Without-RFA | **1023.84** | **18.2027** | **0.6983** |
| Without-CrossAtt | 973.16 | 18.4690 | 0.7314 |

# Outperforming on Complex Motion

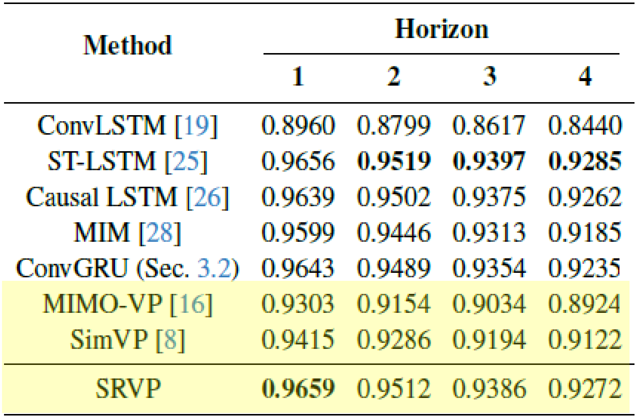
## ✓ SRVP Achieved Superior Performance



| Method | MSE ↓ | PSNR ↑ | SSIM ↑ |
|---|---|---|---|
| ConvLSTM [19] | 936.69 | 17.1198 | 0.5322 |
| ST-LSTM [25] | 386.21 | 21.9665 | 0.7294 |
| Causal LSTM [26] | 408.14 | 21.5750 | 0.7158 |
| MIM [28] | 445.27 | 21.3214 | 0.6947 |
| ConvGRU (Sec. 3.2) | 418.09 | 21.5636 | 0.7115 |
| MIMO-VP [16] | 502.15 | 21.0167 | 0.7276 |
| SimVP [8] | 452.65 | 21.4365 | 0.7179 |
| SRVP | **383.16** | **22.0322** | **0.7360** |

↓ 59.09% in MSE
↑ 29.69% in PSNR
↑ 38.29% in SSIM

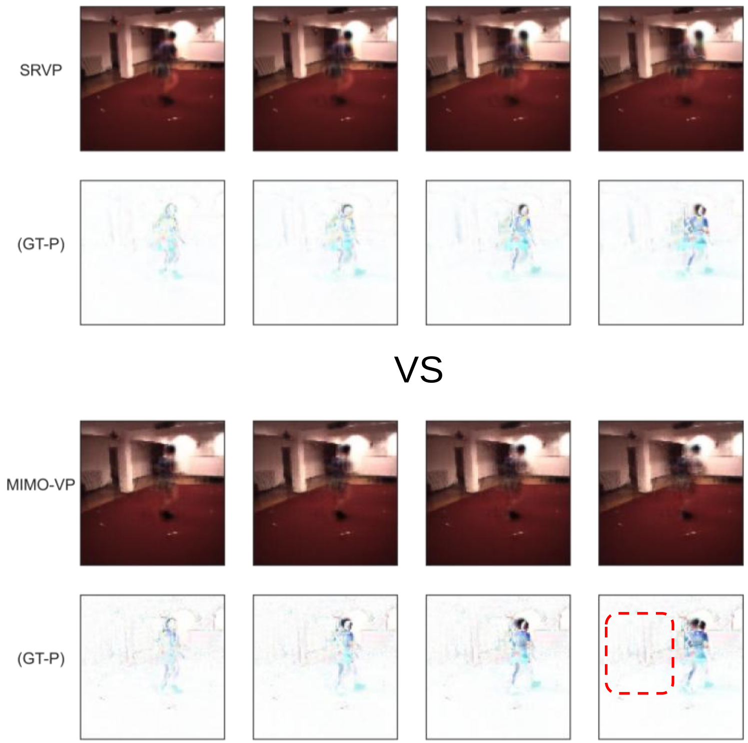
## ✓ Outperform than RNN-free Models



| Method | Horizon | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| ConvLSTM [19] | 0.8960 | 0.8799 | 0.8617 | 0.8440 |
| ST-LSTM [25] | 0.9656 | **0.9519** | **0.9397** | **0.9285** |
| Causal LSTM [26] | 0.9639 | 0.9502 | 0.9375 | 0.9262 |
| MIM [28] | 0.9599 | 0.9446 | 0.9313 | 0.9185 |
| ConvGRU (Sec. 3.2) | 0.9643 | 0.9489 | 0.9354 | 0.9235 |
| MIMO-VP [16] | 0.9303 | 0.9154 | 0.9034 | 0.8924 |
| SimVP [8] | 0.9415 | 0.9286 | 0.9194 | 0.9122 |
| SRVP | **0.9659** | 0.9512 | 0.9386 | 0.9272 |

1.64 to 3.9% in SSIM

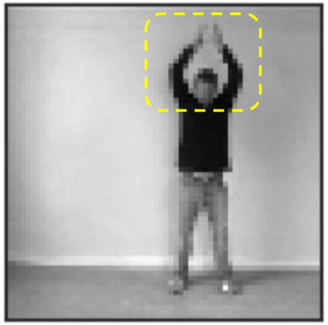
Stronger Robustness and Consistency



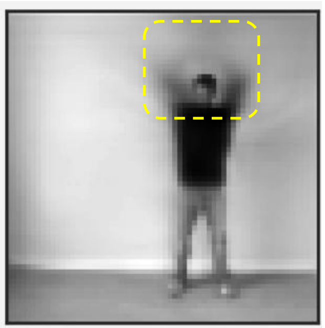
Producing Errors even in Static Regions

# Conclusion

Ground Truth

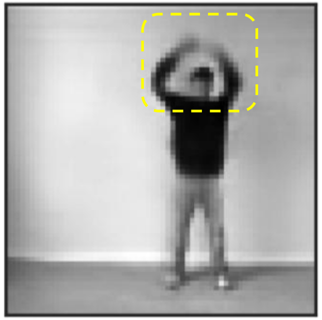


RNN-based



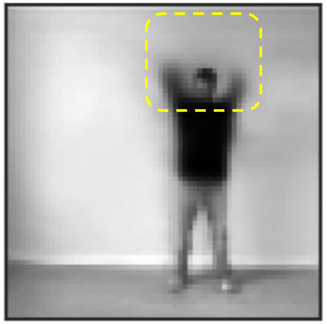
- ➢ Error Accumulation
- ➢ Blurring

RNN-free



- ➢ Temporal Inconsistency (Inaccurate Position)
- ➢ Sensitive to image size changes

SRVP



- ➢ Improving Blurring Problems
- ➢ Consistent Performance
- ➢ Future Work
  → Enhancing Recollection Ability

Thank you for your attention!

https://github.com/yuseonk/SRVP