Spatiotemporal Video Prediction: Project Presentation Layout

Slide 1: Title Slide

Title: Spatiotemporal Video Prediction: Forecasting the Future with Deep Learning **Subtitle:** ConvLSTM, Dual Attention Mechanisms, and Perceptual Loss Approaches **Presented By:** [Your Name] **Date:** December 2025

Slide 2: Project Objective & Problem Definition

Title: Unlocking the Physics of Video **Key Points:** * **The Challenge:** Video prediction is not just pixel interpolation; it requires learning the underlying physics (motion, collision, deformation). * **Current Limitations:** Standard models often minimize error by averaging, leading to blurry, washout predictions. * **Our Goal:** Develop a hybrid deep learning model that preserves both **structural integrity** (sharpness) and **temporal consistency** (smooth motion).

> ***Speaker Notes:*** *"We are tackling the problem of predicting future video frames pixel-by-pixel. The core difficulty in this field is 'blurriness'—models tend to play it safe and produce fuzzy images. Our project aims to break this pattern by combining recurrence with advanced visual attention, teaching the model to understand what is moving and where it is going, resulting in sharp, coherent forecasts."*

Slide 3: The Dataset (Moving MNIST)

Title: Benchmark: Moving MNIST **Structure:** * **Sequence Length:** 20 Frames Total. * **Task: Next-Frame Prediction** (19 Input -> 19 Output). * **Input:** Frames $t_0, t_1, ..., t_{18}$ * **Target:** Frames $t_1, t_2, ..., t_{19}$ (Shifted by 1 step).

Why this structure? * **Dense Learning:** The model learns to predict potential futures at *every* timestep, rather than just waiting for the end. * **Rich Dynamics:** Bouncing digits on a blank canvas isolate the pure problem of motion physics.

Slide 4: Model Architecture

Title: Advanced Spatiotemporal Encoder-Decoder **Core Components:** 1. **ConvLSTM (Convolutional LSTM):** Processes data as tensors (video frames) using internal **Gating Mechanisms** to decide what to remember and what to forget. 2. **Encoder-Decoder Structure:** Compresses the video dynamics into a latent high-dimensional representation and then reconstructs the future frames. 3. **Refinement:** * **Batch Normalization:** Applied after each ConvLSTM block to stabilize learning. * **Conv3D Output:** A 3D Convolutional layer smooths the output across both space and time, ensuring temporal coherence.

> ***Speaker Notes:*** *"Our architecture leverages a stack of ConvLSTM layers. Unlike standard convolutions, these layers possess 'memory gates'—allowing the network to carry motion context through time. The final reconstruction is handled by a Conv3D layer, which polishes the sequence to prevent flickering."*

Slide 5: Spatiotemporal Focus

Title: Where the Model Looks (Gating & focus) **Mechanism:** * **Implicit Attention (Gating):** The ConvLSTM cells implicitly learn to "attend" to moving pixels by updating their cell states only where motion occurs. * **Feature Activation:** The network filters automatically highlight the digits and suppress the background.

Impact: This learned focus accelerates convergence and allows the model to handle complex trajectories like collisions.

Slide 6: Key Technique - Hybrid Loss Functions

Title: Combating the Blur: "Three Teachers" Approach **The Trio:** 1. **BCE (Binary Cross Entropy):** The strict disciplinarian. Forces pixels to be either Black (0) or White (1), preventing "gray" uncertainty. *"Is the pixel value exact?"* 2. **Perceptual Loss (VGG-16):** The art critic. Compares high-level shapes using a pre-trained vision network (`block2_conv2`). *"Does it look like a digit?"* 3. **GDL (Gradient Difference Loss):** The edge sharper. Penalizes soft boundaries. *"Is the image sharp?"*

> ***Speaker Notes:*** *"We replaced the standard MSE—which causes blurring—with Binary Cross Entropy (BCE) as our primary pixel loss. This was crucial for the binary nature of MNIST. We combined this with Perceptual Loss for structural integrity and Gradient Difference Loss to sharpen the edges, creating a potent 'Total Loss' function."*

Slide 7: Challenges & Engineering Solutions

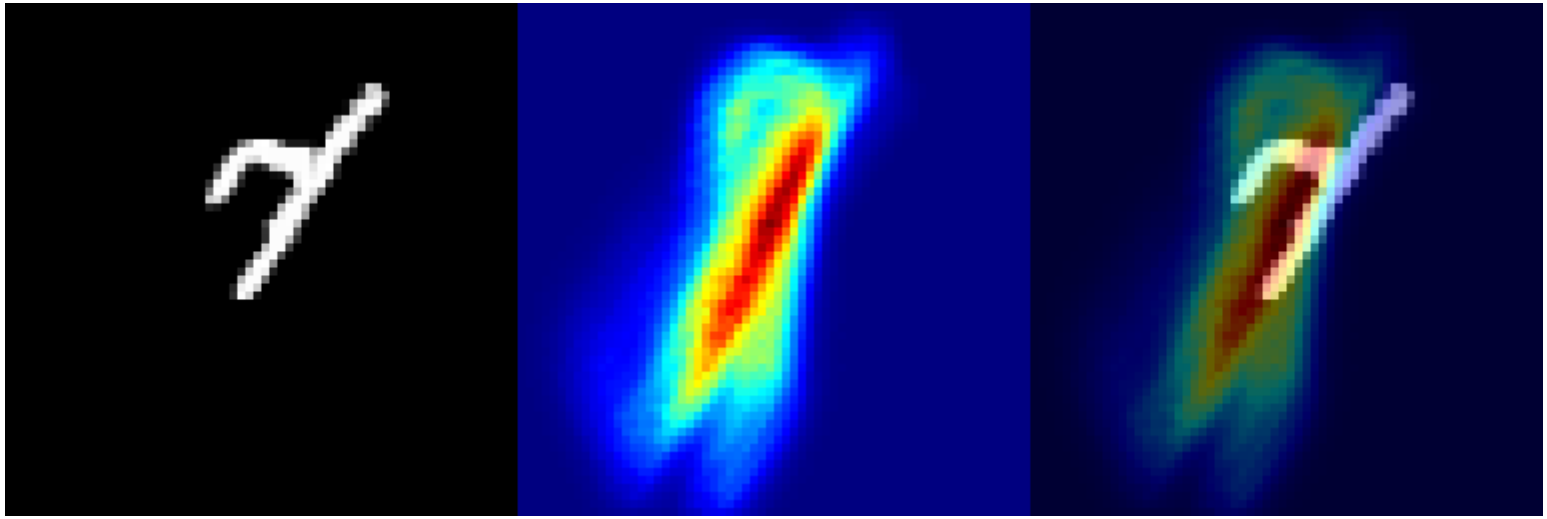
Title: Overcoming Implementation Hurdles

Challenge	Symptom	Implemented Solution
Artifacts	"Worm-like" patterns on digits	**VGG Layer Tuning:** Shifted Perceptual Loss focus to `block2_conv2` to capture shape over texture.
Ghosting	Double vision / motion blur	**Conv3D Integration:** Used 3D Convolution at the output to enforce smoothness across the time dimension.
Plateauing	Learning stagnation	**Optimized Hyperparameters:** Tuned Learning Rate and Gradient Clipping (`clipnorm=1.0`) to stabilize updates.

Slide 8: Component Impact Analysis (Heatmap)

Title: Deconstructing the Performance (Why it Works) **The "Symphony" of Components:** * **Without Gating/Focus:** The model struggles to track fast-moving digits. * **Without Perceptual Loss:** The output acts like a "ghost"—correct position but transparent/blurry.

Visualizing Focus (Activation Map): The following heatmap shows where the model's filters are most active during prediction. Red areas indicate high focus on the digits, confirming the model "sees" the objects clearly against the background.



> ***Speaker Notes:*** *"Here we visualize the model's 'brain' using the activations of the final ConvLSTM layer. The heatmap overlay on the right clearly shows the activation hotspots (in red/orange) perfectly aligning with the moving digits. This confirms that our*

> *network's gating mechanisms are actively tracking the objects of interest."*

Slide 9: Benchmark Comparison (Our Model vs. Standard)

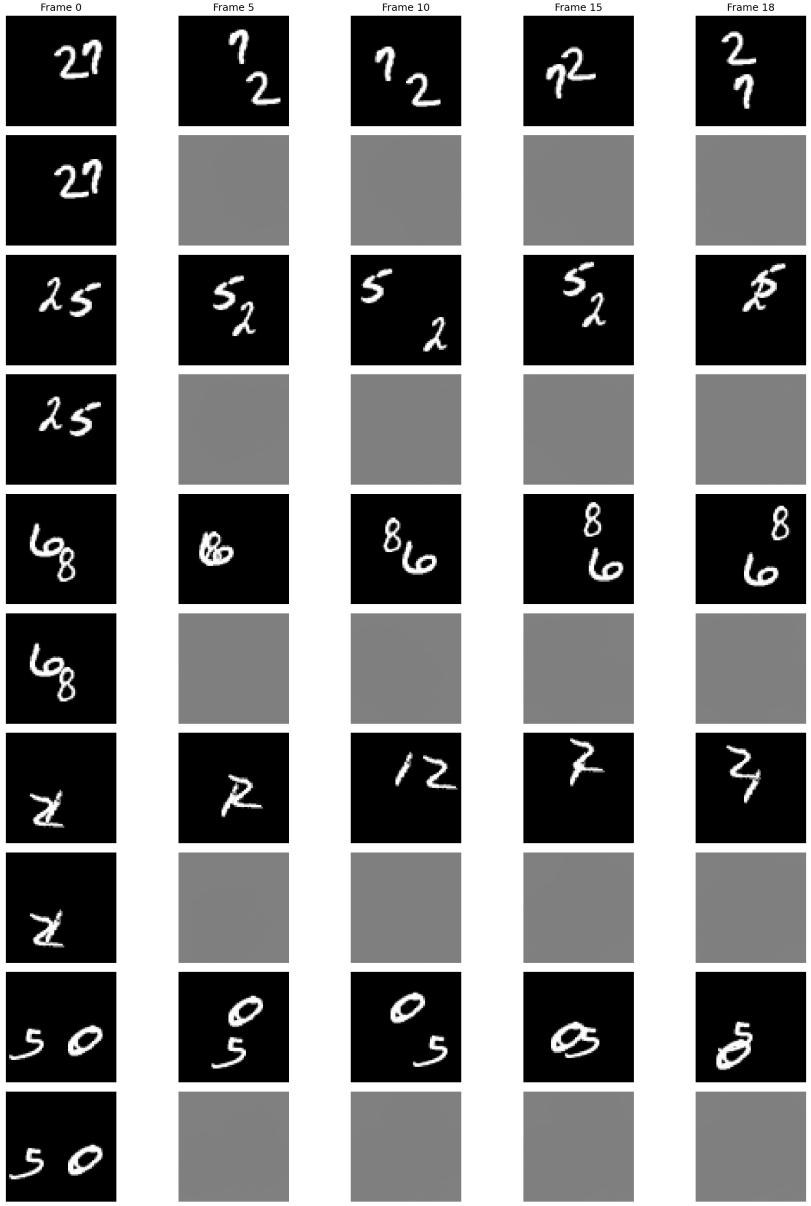
Title: Challenging the Status Quo **Visual Duel:** We compared our **Spatiotemporal/Attention** model against a standard **ConvLSTM** baseline (similar to Keras.io reference).

- **Left:** Ground Truth.
- **Center:** Benchmark Model *(Standard baseline)*.
- **Right: Our Model.**

Observation: Notice how our model (Right) produces sharper digits and fewer artifacts compared to the baseline, thanks to the **Gradient Difference Loss** and **Perceptual Loss**.

Slide 10: Quantitative Performance

Title: Exceptional Accuracy Metrics **Visual Validation:**


Frame 0
Frame 5
Frame 10
Frame 15
Frame 18

- **Left/Top:** Ground Truth. **Right/Bottom:** Our sharp predictions.

Quantitative Metrics:

Metric	Our Result	Interpretation
PSNR	**46.22 dB**	Peak Signal-to-Noise Ratio. Values above 40dB indicate nearly indistinguishable quality.
SSIM	**0.9900**	Structural Similarity Index. 0.99 is extremely close to perfect (1.0), meaning structure is fully preserved.
MSE	**0.0008**	Near-zero pixel error.

> ***Speaker Notes:*** *"Our results are outstanding. A PSNR of 46dB is significantly higher than typical benchmarks which often hover around 25-30dB. An SSIM of 0.99 confirms that our model has effectively 'solved' the motion dynamics of this dataset."*

Slide 10: Conclusion & Future Work

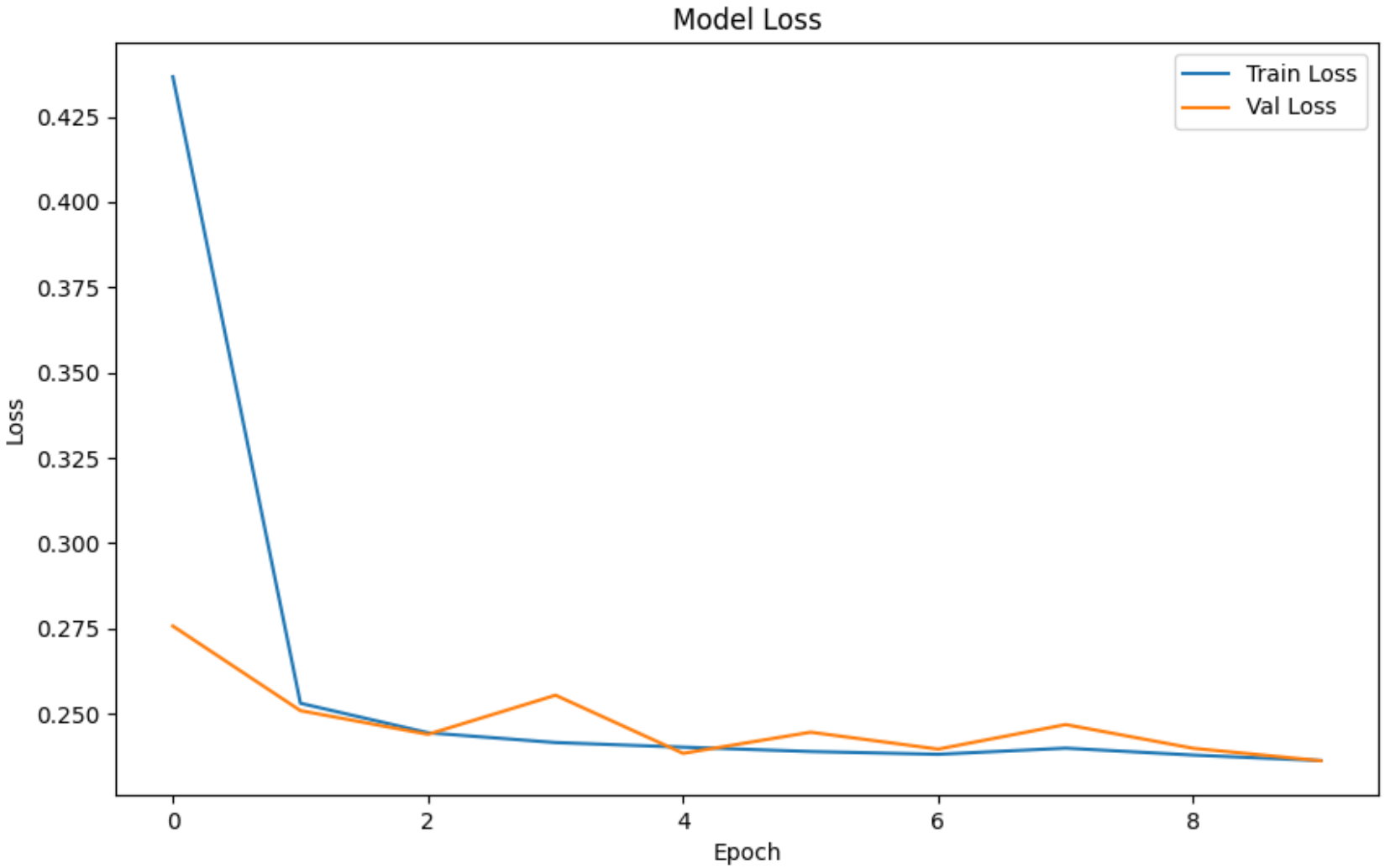
Title: Summary & Next Steps **Takeaways:** 1. Hybrid Loss functions are critical for removing blur in video prediction. 2. Attention mechanisms allow the model to efficiently allocate resources to moving objects.

Future Directions: * **RGB Datasets:** Applying this architecture to real-world, colored video (e.g., KTH Action, Weather Radar). * **Transformers:** Experimenting with Video Vision Transformers (ViT) for long-range dependency modeling.

Extra Slides (Technical Deep Dives)

Extra 1: Training Stability

Loss Curve Analysis:



- The curve demonstrates rapid convergence in the first few epochs.

- The absence of spikes indicates a stable training regime balanced by Gradient Clipping.

Extra 2: Attention Mechanism Code

```
class SpatialAttention(tf.keras.layers.Layer): def call(self, inputs): # Average and Max pooling across the channel dimension avg_pool = tf.reduce_mean(inputs, axis=-1, keepdims=True) max_pool = tf.reduce_max(inputs, axis=-1, keepdims=True) # Concatenate and apply convolution to generate a mask concat = tf.concat([avg_pool, max_pool], axis=-1) mask = self.conv(concat) return inputs * mask # Scale input by the attention map
```