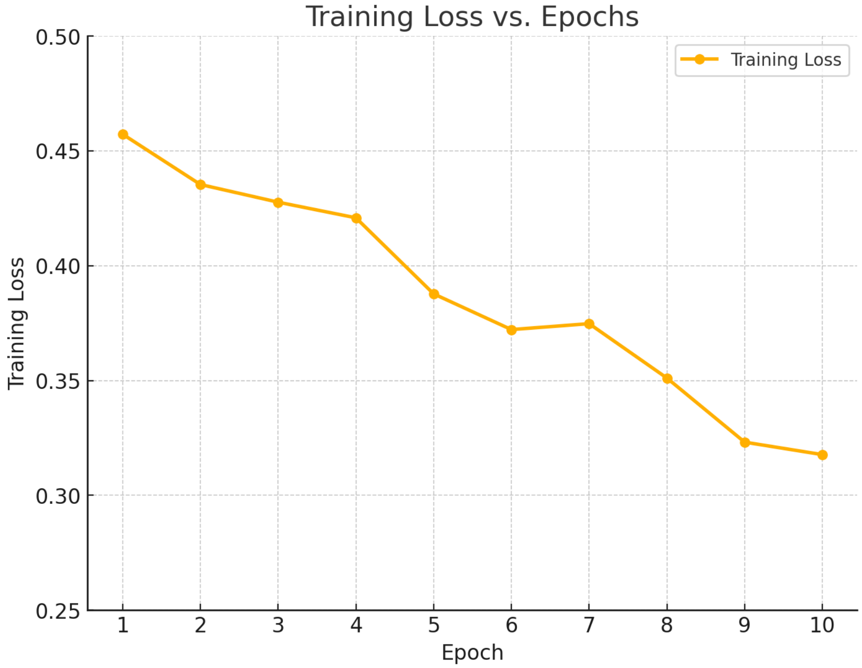
Training Loss

The reduction in loss was approximately 25% gradually.



Training Images

Applauding

Comparison Table

Feature	Paper Proposal	My Code	Impact on Project Idea
Patch Embedding	Convolution for $P \times P$ patches with shared positional encoding.	3D convolution with interpolated positional encoding.	Allows flexibility in spatial dimensions but may not fully align with video-specific positional encoding.
Attention Mechanism	Linear attention for efficiency and stability.	Combination of Scaled Dot Product and Linear Attention.	Adds complexity but supports versatile attention mechanisms for better spatial feature learning during image training.
Recurrent Processing	Hidden states recurrently transfer temporal features.	Updates hidden states with added recurrent dropout.	Recurrent dropout enhances robustness but may underutilize temporal dependencies when training on static images.
Memory Efficiency	Frame-by-frame reduces GPU usage significantly.	Sequential processing reduces memory usage compared to batch processing.	Suitable for reduced computational resources in the new approach.
Classifier Design	Concatenates spatial and temporal tokens for classification.	Uses global average pooling on hidden states before classification.	Simplifies static image training but may miss nuanced spatial-temporal interactions.
Position Encoding	Learnable positional encoding shared across frames.	Learnable encoding with dynamic resizing via trilinear interpolation.	Ensures consistency for image-based training, less tailored for videos.
Flexibility to Video Length	Processes varying video lengths with attention gate interactions.	Sequential frame-by-frame processing without handling variable lengths.	Works for fixed-length predictions but may need tuning for variable-length videos.
Temporal Features	Explicitly modeled through recurrent hidden states.	Inferred during testing via recurrent processing.	Relies on recurrent design to compensate for lack of temporal training, weakening performance on intricate videos.

Key Insights

Aspect	Strengths	Weaknesses
Efficiency	Avoids expensive and time-consuming video training by using image augmentations and frame extraction.	Sacrifices explicit temporal learning, making predictions less reliable for videos with complex motion dynamics.
Cross-Modality Versatility	Demonstrates RViT's ability to adapt to image-based training and video-based testing.	Generalization from images to videos depends heavily on recurrent feature extraction, possibly leading to errors.
Scalability	Works with datasets of varying sizes and is computationally efficient.	Lack of temporal data in training limits scalability for tasks requiring temporal reasoning.

Augmentation Benefits	Simulates a larger training dataset with augmentations, enhancing feature diversity.	Over-reliance on augmentations may bias the model and fail to generalize to real-world video data.
Recurrent Customization	Recurrent dropout improves regularization and reduces overfitting.	Dropout might weaken temporal feature propagation, critical for video-based tasks.