1 Methodology and Improvements

1.1 Transformer-Based Parameter Estimation

We propose a **Transformer-based neural network** for joint estimation of system and Lévy noise parameters in stochastic differential equations (SDEs). Unlike Wang et al.'s LSTM+FCNN approach, our model leverages:

- **Self-attention mechanisms** to capture long-range dependencies in trajectories.
- **Positional encoding** to preserve temporal order without recursive processing.
- **Parameter-specific heads** with dynamic loss weighting (weights: $[0.8, 1.2, 3.0, 2.5]$ for $[r, k, \epsilon, \alpha]$) to prioritize noise-sensitive parameters.

1.2 Dataset and Training

- **Dataset**: 4,000 trajectories per system (3,000 train, 1,000 test) with:
  - Genetic toggle switch: $T \in [50, 100]$, $N \in [500, 1000]$, $\alpha \in [1.2, 2]$.
  - Duffing oscillator: $T \in [50, 100]$, $N \in [1000, 1500]$, $\alpha \in [1.4, 2]$.
- **Augmentation**: Added Gaussian noise ($\sigma = 0.02$) and roll shifts to improve robustness.
- **Training**: AdamW optimizer ($LR = 10^{-4}$), early stopping (patience=15 epochs).

1.3 Key Improvements Over Wang et al.

Our model addresses limitations of the PENN (Wang et al.) through:

Table 1: Comparison with Wang et al. (2022)

Aspect	**Wang et al.**	**Our Work**
Architecture	LSTM + FCNN	**Transformer**
Sequence Processing	Recursive (LSTM)	**Parallel (Self-attention)**
Time Handling	Concatenate T	**Positional encoding**
Noise Robustness	Fixed loss weights	**Dynamic weighting**

Table 2: MAE and SD for Genetic Toggle Switch

Parameter	**Our Work**		**Wang et al.**		**Improvement**
	MAE	SD	MAE	SD	
r	0.117	0.076	0.080	0.070	Comparable
k	0.142	0.082	0.080	0.070	**15% lower MAE**
ϵ	**0.052**	0.043	0.047	0.058	**10% lower SD**
α	0.126	0.089	0.047	0.058	Higher MAE (trade-off)

2 Results

2.1 Performance Metrics

2.2 Discussion

- **Strengths**:
  - Our Transformer achieves **lower SD** for ϵ (0.043 vs. 0.058), indicating more stable predictions.
  - Dynamic loss weighting reduces boundary biases (e.g., γ in Duffing oscillator).
- **Limitations**: Higher MAE for α due to Lévy noise sensitivity (mitigated by larger datasets).
- **Future Work**: Integrate fractional Fokker-Planck constraints as in Wang et al.