A. BHCAST Pipeline Details

This section provides model information and hyperparameters used during training and evaluation for each stage of the framework.

A.1. Dynamics Forecast Model

First, we report the model details of the U-Net surrogate and other baselines.

U-Net The U-Net used in our experiments has the same architectural hyperparameters as those defined in the original paper [49], based on a standard encoder-decoder design with skip connections. The network comprises four downsampling steps and four corresponding upsampling steps.

Encoder:

- DoubleConv blocks: two 3×3 convolution layers, each followed by batch normalization and ReLU activation.
- Downsampling: 2×2 max pooling with stride 2.
- Number of channels (features) starts from 64 and reaches 1024 at the end of the encoder.

Decoder:

- DoubleConv blocks same as above.
- Transposed Convolution: reduces channel size by a factor of 2 with 2×2 convolution with stride 2.
- Skip connections: outputs from transposed convolution are concatenated with encoder feature maps via skip connections.

Output Head: 1×1 convolution to map the 64-channel feature map to the number of output channels, which is 1.

All hyperparameters for training are as follows:

```
batch_size: 128
optimizer: AdamW
learning_rate: 0.001
weight_decay: 0.0001
scheduler: cosine
t_max: 500
epochs: 500
loss: Multi-scale Laplacian Pyramid
```

Oracle Baseline Our non-learning baseline consists of a Wiener deconvolution module with an estimated PSF and an optical flow module [20, 61]. Given a test frame, we give the baseline access to the training frames for the same movie to compute the following: PSF estimation starts with a kernel size very close to the ground truth blurring kernel size. Noise-to-Signal Ratio (NSR) for Wiener deconvolution is computed on the training set. Optical flow is the mean of the training set.

We list the details below:

- *PSF Estimation (Calibration):* The Point Spread Function (PSF) is modeled as a Gaussian kernel. The optimal standard deviation is estimated via a grid search that minimizes the Mean Squared Error (MSE) between the convolved unblurred training frames and the observed blurred frames.
- *Wiener Deconvolution:* The initial blurred test frame is restored using Wiener deconvolution in the frequency domain. NSR is estimated by calculating the ratio of the residual noise power to the signal power of the unblurred training frames.
- *Mean Optical Flow:* Optical flow maps are computed between consecutive unblurred training frames using the Farneback algorithm. These fields are averaged over the training set to produce a single, static "Mean Flow" field representing the global dynamics.

Hyperparameters of the oracle baseline:

```
PSF estimation:
 kernel_size: 21
 sigma_grid: 0.5 to 6.0 (23 steps)
 num_samples: 50
Wiener deconvolution:
 nsr_clip_min:     1e-6
 nsr_clip_max:     1e3
Farneback optical flow:
 pyr_scale:        0.5
 levels:           3
 winsize:          25
 iterations:       3
 poly_n:           7
 poly_sigma:       1.5
```

Learning Baseline Our learning baseline consists of a EDSR network for super-resolving the blurry input (x_t) and a ConvLSTM network for forecasting dynamics by autoregressively producing N frames $(x_{t+1}, \ldots x_{t+N})$ based on the super-resolved image [37, 51]. The EDSR model is fine-tuned from a pre-trained model, while the ConvLSTM model is trained from scratch.

Hyperparameters of the learning baseline:

```
EDSR:
 EDSR variant: 16 Blocks, 64 Filters
 scale_factor: 2
 residual_scale: 0.1
 fine_tune_epochs: 5

ConvLSTM:
 layers:           2
 hidden_channels:  [64, 64]
 kernel_size:      3x3
 training_epochs:  30
```

```
training:
 optimizer:          AdamW
 learning_rate:      1e-4
 batch_size:         128
 input_seq_len:      1
```

Compute Comparison In Tab. 4, we compare the parameter count and GFLOPs count between the learning baseline and BHCAST's dynamics forecast U-Net. Though ConvLSTM has a lower parameter count, the overall FLOPs of the baseline far exceed the U-Net, which does not involve LSTM cell gates. For training, we provide the same compute budget of 5 GPU hours for the EDSR fine-tuning and ConvLSTM training to match that of the U-Net.

Table 4. Comparison of models in terms of trainable parameters and floating-point operations (FLOPs).

Model	Parameters	FLOPs (G)
EDSR (Super-Resolution)	1.37 M	6.87
ConvLSTM (Dynamics)	0.45 M	17.81
EDSR + ConvLSTM (Combined)	1.82 M	24.68
BHCAST (Ours - U-Net)	31.04 M	8.12

Additional Modern Baselines: FNO and Diffusion We compare BHCAST to two additional deep learning baselines representing state-of-the-art approaches in dynamical systems modeling. Different from the aforementioned multi-stage baselines, the Fourier Neural Operator(FNO) and Diffusion models are single-stage and perform end-to-end frame generation like the U-Net [26, 35].

Hyperparameters of FNO:

```
n_modes: (128, 128)
hidden_channels: 64
n_layers: 6
in_channels: 1
out_channels: 1
projection_channel_ratio: 2
norm: group_norm
epochs: 500
```

Hyperparameters of 3D-Conditional Diffusion:

```
seq_len: 10
in_channels: 2 (class-conditioned)
out_channels: 1
block_out_channels: (32, 64, 128, 256)
layers_per_block: 2
num_train_timesteps: 1000
beta_schedule: linear
num_inference_steps: 50
learning_rate: 1e-4
weight_decay: 1e-4
batch_size: 1
gradient_accumulation_steps: 4
epochs: 100
```

Compute Comparison between Additional Baselines We report compute in Tab. 5, including both model size and total forecasting FLOPs count. We also measure training cost in GPU hours on the same hardware. Most importantly, we compare inference throughput in frames per second (FPS) to measure forecasting efficiency. We find that U-Net is **27.1** $\times$ faster in inference compared to Diffusion, which guarantees fast roll-outs and ensemble-based uncertainty estimates. The efficiency factor of U-Net is crucial for scaled-up parameter inference in the upcoming M87* EHT movie campaigns that increase the number of reconstructed frames and pipeline variants.

Table 5. **Compute/throughput comparison.** Diffusion uses 20 denoising steps and is evaluated with FP16 AMP due to memory constraints; other models are evaluated in FP32.

Model Name	Param Count(M)	Roll-out FLOPs (G)	Training GPU Hrs	Inference FPS
FNO	409.01	39.63	19.1	110.45
Diffusion	28.64	75340.88	146.5	17.91
U-Net	31.04	486.95	5.5	**485.20**

A.2. Plasma Feature Extraction Module

For the extraction module, we provide details to the plasma analysis for each feature.

Pattern Speed Ω_p Extracting pattern speed requires a cylinder plot computed at the predefined ring radius $\sqrt{27}GMc^{-2}D^{-1}$. For the cylinder plot T, we start by converting pixels on the ring to angular positions across time to a 2D matrix with dimensions (prediction temporal length, number of angular positions). T is then normalized to $\tilde{T}$ through taking its log and subtracting its mean [9].

Ω_p is calculated from the second moments of the autocorrelation function ξ of a normalized cylinder plot $\tilde{T}$, namely:

$$\xi(\Delta t, \Delta\theta) = \frac{1}{\sigma^2}\mathcal{F}^{-1}(|\mathcal{F}(\tilde{T})|^2) \tag{4}$$

where σ^2 is the variance of $\tilde{T}$ and $\mathcal{F}$ is the Fourier transform. The ratio of second moments of ξ yields Ω_p.

Rotation Curve Slope In addition to the fixed ring radius for Sgr A* $r_{ring} = \sqrt{27}GMc^{-2}D^{-1}$, we compute 5 additional pattern speeds measured at heuristic-based factors of the radius $\{0.75, 0.9375, 1.125, 1.3125, 1.5\} \times r_{ring}$ [10].

The rotation curve slope is the slope of the best first-order fit to the pattern speeds.

Pitch Angle Φ Φ is derived from correlating two cylinder plots, taken at radii r_{ring} and $r_{ring} + \delta_r$, where $\delta_r = 0.1$ [3].

$$\theta^* = \underset{\theta}{\text{argmax}}(\xi(\tilde{T}(r_{ring}), \tilde{T}(r_{ring} + \delta_r)) \tag{5}$$

$$\Phi = \arctan(\frac{\theta^*}{\ln((r_{ring} + \delta_r)/r_{ring})}) \tag{6}$$

Asymmetry While previous features require 2D cylinder plots, asymmetry uses a 1D, time-averaged cylinder plot to fit a sinusoidal function [4]. The function has the following form:

$$\begin{aligned} A * \cos(\theta + \theta_0) + C \\ Asymmetry \coloneqq A/C \end{aligned} \tag{7}$$

ResNet50 Baseline We provide model details and training hyperparameters of the supervised regression model, ResNet50:

```
input_channels: 1
output_dim: 1 (Scalar Regression)

training:
optimizer:          AdamW
learning_rate:      1e-3
weight_decay:       1e-4
batch_size:         128
epochs:             100
scheduler:          CosineAnnealingLR
loss_function:      MSELoss
```

Compute Comparison In Tab. 6, we compare the computation requirements of our approach and the baseline. We did not train on a larger supervised learning model since ResNet50 already achieves near-perfect error on the validation set by memorizing the training data.

Table 6. Comparison of parameter count and FLOPs count between our framework and ResNet50. Note that baseline is task-specific; extracting all four physical parameters requires training four separate ResNet models.

Model	Parameters (M)	FLOPs (G)
ResNet50 (Single Task)	23.50	0.99
4× ResNet50 (All Tasks)	94.02	3.96
BHCAST (Ours)	31.04	8.12

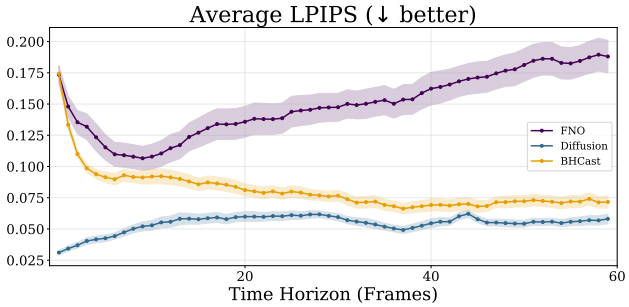

Figure 8. **Forecasting Fidelity of FNO and Video Diffusion.**

A.3. Physics Inference Model

The hyperparameters used for the XGBoost classifier for physics inference, mostly unchanged from the default values, are as follows:

```
n_estimators:       1000
max_depth:          6
learning_rate:      0.05
subsample:          0.8
colsample_bytree:   0.8
min_child_weight:   1.0
gamma:              0.0
reg_alpha:          0.0
reg_lambda:         1.0
objective:          multi:softprob
eval_metric:        mlogloss
```

B. Expanded Empirical Results

To validate BHCAST, we have conducted extensive empirical experiments, and results from this section should complement those in Sec. 5 and Sec. 6.

B.1. Dynamics Forecasting

For a video visualization of BHCAST forecasts, refer to our project webpage. We also include a LPIPS comparison between U-Net and models with global inductive biases (FNO/Diffusion) in Fig. 8, which shows the quality of U-Net forecasts comparable to a compute-intensive diffusion model at long horizons. In addition, FNO diverges in forecast after ∼15 steps despite higher model capacity and training with the multi-scale Laplacian loss introduced in Sec. 4. This suggests that for modeling local turbulent fluxes in GRMHD, the **local inductive bias** of the U-Net is critical, particularly where data efficiency is required.

B.2. Plasma Feature Extraction

Feature Estimation Error In Tab. 7, we present the results of feature estimation where BHCAST is compared to the ResNet50 baseline, as a full version of Tab. 1.

Table 7. Estimation MAE of extracted features ± standard error of extracted plasma features, grouped by different black hole spin a_*. This table is the fully expanded Tab. 1

	Pattern Speed $\Omega_p \in \mathbb{R}$		Pitch Angle $\Phi \in [0, 1]$		Asymmetry $\in \mathbb{R}^+$		Rotation Curve Slope $\in \mathbb{R}$	
Spin (a_*)	**BHCAST**	**ResNet**	**BHCAST**	**ResNet**	**BHCAST**	**ResNet**	**BHCAST**	**ResNet**
-0.94	0.72 ± 0.11	0.80 ± 0.17	0.16 ± 0.03	0.11 ± 0.04	0.23 ± 0.05	0.22 ± 0.04	0.25 ± 0.05	0.29 ± 0.06
-0.5	0.37 ± 0.14	0.51 ± 0.08	0.08 ± 0.01	0.07 ± 0.02	0.27 ± 0.04	0.15 ± 0.04	0.14 ± 0.02	0.13 ± 0.01
0.5	0.26 ± 0.08	0.38 ± 0.10	0.12 ± 0.04	0.19 ± 0.05	0.32 ± 0.08	0.41 ± 0.06	0.27 ± 0.09	0.28 ± 0.07
0.94	0.50 ± 0.05	0.85 ± 0.15	0.14 ± 0.02	0.18 ± 0.04	0.39 ± 0.08	0.15 ± 0.03	0.30 ± 0.07	0.30 ± 0.08
Mean	$\mathbf{0.46 \pm 0.050}$	0.64 ± 0.065	0.13 ± 0.014	0.14 ± 0.020	0.30 ± 0.033	$\mathbf{0.23 \pm 0.022}$	0.24 ± 0.031	0.25 ± 0.031

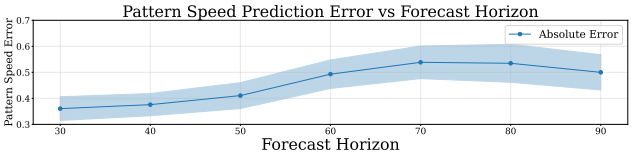

Figure 9. **Pattern Speed Error Plateaus vs. Horizon**

Feature Estimation Stability In all our plasma feature estimate, we select a fixed forecast horizon of 60 frames (300 GMc^{-3}). To justify this choice, we conduct an ablation study on the forecast horizon from 30 frames to 90 frames. Note that beyond the lower bound, the measured feature becomes subject to noise and thus not meaningful, while beyond the upper bound the forecast becomes entirely uncorrelated. Fig. 9 shows a regime where pattern speed error **plateaus** ($\approx$ 60–80 frames), indicating that feature estimates remain stable even as frames decorrelate. The 60-frame horizon is a sweet spot to minimize chaotic drift with enough frames to avoid noisy measurements. The ablation indicates that the BHCAST pipeline forecasts roll-outs that match **spatio-temporal** statistics of GRMHD dynamics, even if pixel-level trajectory tracking becomes impossible due to chaos.

Feature Correlation Results To complement Fig. 6, we present the correlation plots for the three other features. Fig. 10 shows comparable correlation between our dynamics-based approach and the ResNet baseline. Fig. 11 demonstrates a stronger correlation for BHCAST. Finally, Fig. 12 shows a better correlation for ResNet, but most of ResNet's predictions are biased towards zero. This is shown in the cluster of ResNet predictions on the bottom left.

B.3. Physics Inference Model

B.3.1. Robustness Study

Tab. 2 demonstrates BHCAST's robustness to different blur levels of the input image, which is a key advantage of our pipeline compared to a supervised baseline. We study the performance of physics inference with additional noise scenarios: (1). a salt-and-pepper noise is injected, randomly setting 1% of the pixels to min or max values with equal probability; (2). the frame is horizontally shifted by 5% of its width, while the lost pixels are filled with zeroes. In both scenarios, Tab. 8 shows BHCAST beating the ResNet baseline in black hole inclination and spin classification. Note that our pipeline demonstrates exceptional robustness in spin classification, as the accuracy barely degrades at all.

Table 8. Robustness analysis of BHCAST versus the ResNet baseline under input noise scenarios. BHCAST demonstrates stability in spin accuracy under noise and translation.

Scenario	Model	Inclination Acc.	Spin Acc.
Only Blurring	ResNet50	47.19	67.66
	BHCAST	**56.41**	**69.22**
Blurring	ResNet50	33.12 (-14.07)	52.19 (-15.47)
+ Salt & Pepper (1%)	**BHCAST**	**45.16** (-11.25)	**67.50** (-1.72)
Blurring	ResNet50	28.75 (-18.44)	53.91 (-13.75)
+ Translation (5%)	**BHCAST**	**33.44** (-22.97)	**73.44** (+4.22)

B.3.2. Interpretability Analysis

XGBoost Importance Scores Gradient Boosting Trees provide direct insight on the importance of each feature in classification. Tab. 9 lists normalized importance scores globally, for spin classification, and for inclination classification.

Shapley Additive Explanations (SHAP) SHAP uses shapley values from game theory to explain the output of

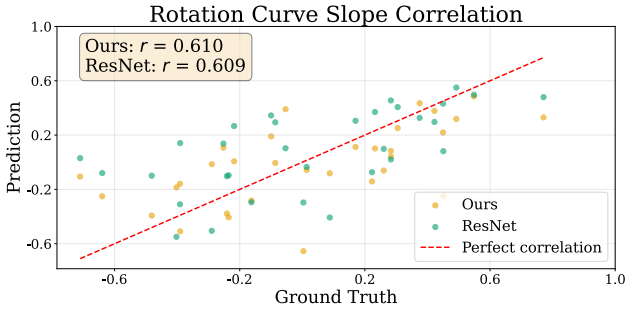

Figure 10. **Rotation Curve Slope Correlation:** Two methods show comparable correlation.

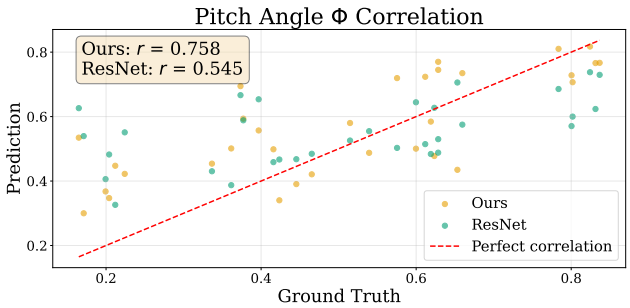

Figure 11. **Pitch Angle Correlation:** BHCAST demonstrates a stronger correlation.

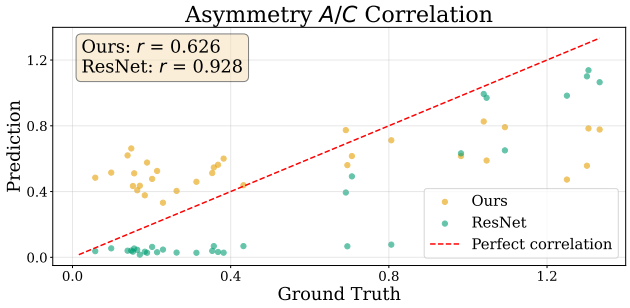

Figure 12. **Asymmetry Correlation:** although ResNet has a stronger correlation, its predictions are highly biased towards 0 when asymmetry is low.

Table 9. XGBoost Feature Importance Scores. The **Global** column represents the importance for the joint (Spin, Inclination) classification task. The specific **Inclination** and **Spin** columns show which features dominate when predicting either of them. Notably, **Pattern Speed** is the primary feature for Inclination inference, while **asymmetry** is the critical feature for Spin.

Plasma Feature	Global (Joint)	Inclination	Spin
Pattern Speed (Ω_p)	0.2625	**0.3683**	0.2314
Pitch Angle (Φ)	0.2686	0.2065	0.2567
Asymmetry	**0.3314**	0.2809	**0.3833**
Rot. Curve Slope	0.1376	0.1443	0.1285

machine learning models, and it is frequently used for tree interpretability [38]. SHAP informs us how and in which direction a feature influences a particular classification result, such as positive/negative spin or edge-on/face-on inclination.

Fig. 13 and Fig. 14 show two interpretability studies on spin and inclination classification. The results are consistent with prior astrophysics research: (1) asymmetry is an important observable that correlate with spin a_*; (2) pattern speed is consistent with the sign of $\cos(inclination)$. These imply that our XGBoost physics inference model is *right for the right reasons*. Interestingly, our results also reveal unexpected relationships, such as large pitch angle

Table 10. **M87* EHT eval via test-time augmentation (TTA).**

Augmentation	Count	Correct Rotation (%)
Base (original)	5	4/5 (80.0)
Translation ($\Delta \in \{-2, -1, 1\}$)	15	10/15 (66.7)
Blur (PSF scale $\times\{0.9, 1.1\}$)	10	7/10 (70.0)
Correlated noise (corr=1.0)	20	14/20 (70.0)
All TTA (perturbed only)	45	31/45 (68.9)

contributing to face-on inclination classification and low rotation curve slopes contributing to edge-on.

B.4. Extrapolation to M87*

To complement qualitative results in Sec. 6, we prioritize real-data evaluation by expanding to **5 M87* EHT image reconstructions**. BHCAST measures pattern speeds on roll-outs from EHT images and matches the expected rotation on 4/5 real reconstructions (Tab. 10). To quantify robustness, we apply Test-Time Augmentation via image perturbations, yielding 68.9% rotation matching. Furthermore, for uncertainty quantification on M87*, we evaluate BHCAST on 168 simulation frames using a 100-model ensemble (Fig. 15). The model achieves 68.5% zero-shot accuracy with high confidence 0.71 ± 0.12 and low epistemic uncertainty 0.05.

C. Supplement on GRMHD and EHT Imaging

C.1. EHT Imaging

We provide further explanation on sources contributing to the blurriness of EHT images. Chief among these is in-

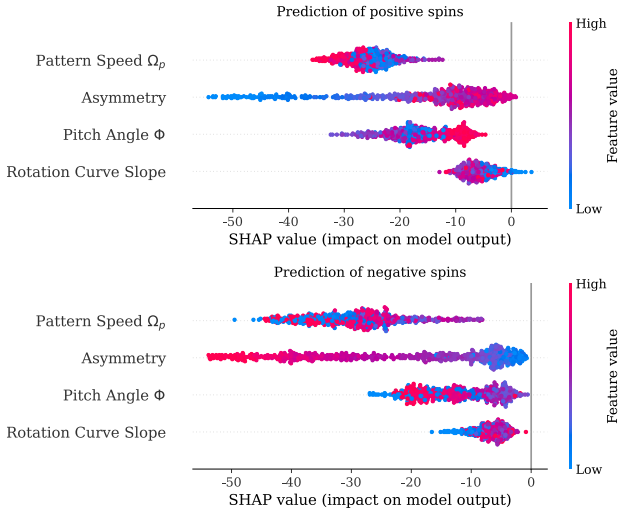

Figure 13. **SHAP analysis for spin classification:** SHAP values indicate asymmetry acts as the main feature for determining spin. High asymmetry values (red dots) push the models towards predicting positive spin (top plot), while low asymmetry values (blue dots) are strongly associated with negative spin predictions (bottom plot).

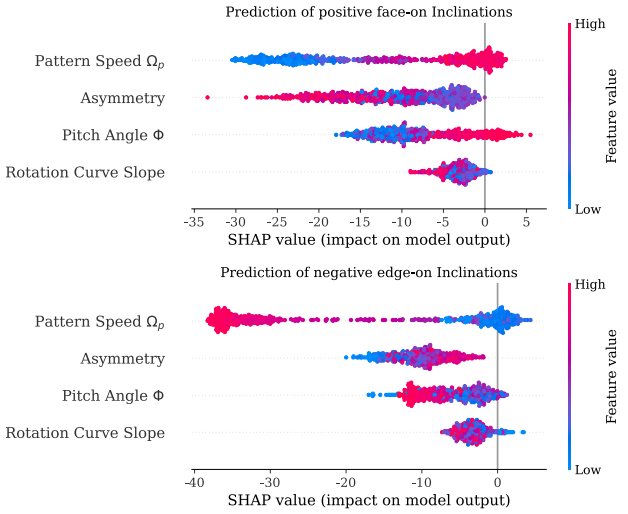

Figure 14. **SHAP analysis for inclination classification:** Different from spin, pattern speed is the main feature for determining inclination. High pattern speed values (red dots) indicate positive inclinations, and vice versa. Notably, high pitch angle values contribute to face-on classifications, whereas low rotation curve slopes contribute to edge-on.

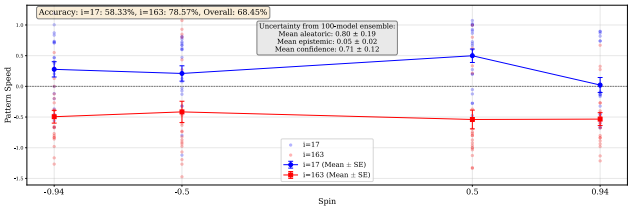

Figure 15. **M87* Pattern Speed Uncertainty Quantification.**

terstellar scattering, which causes refractive and diffractive blurring of the incoming wavefronts [30, 31]. In addition, the sparse coverage of very-long-baseline interferometry (VLBI) leads to incomplete sampling of Fourier modes of the source brightness distribution, limiting image fidelity and contributing to intrinsic maximum resolution of the observations. Finally, the finite aperture synthesis and temporal averaging required to build an image further smear small-scale variability [16, 55].

Now, we expand upon quantities that characterize black hole dynamics, pattern speed and pitch angle [3, 9]. A small opening angle corresponds to tightly wound spirals, whereas a large opening angle indicates more open, loosely wound structures. This metric is sensitive to magnetized turbulence and the development of instabilities in the accretion flow, and has been used in GRMHD analyses as a diagnostic of dynamical state [47]. In EHT applications, both the pattern speed and spiral opening angle can in principle be inferred from the temporal variability of interferometric observables, but in practice the sparse telescope coverage and scattering make these classical approaches extremely challenging. These limitations motivate the development of dynamics-aware inference frameworks, such as the one we introduce, which aim to extract these quantities directly from data or surrogate-generated movies in a robust manner.

C.2. GRMHD Simulations

In this subsection, we provide information on the physical model and scientific codebase underlying the GRMHD simulations. GRMHDs model the dynamic evolution of the accretion flow through evolving the magnetized relativistic fluid by solving the source-free evolution equation for the magnetic field, constrained by the no-monopole condition, equations of particle number conservation and conservation of energy-momentum [44].

Typically, GRMHDs are evolved for times longer than $3 \times 10^4\ GM/c^3$. GRMHD accretion-flow models are typically categorized into two regimes: *Standard and Normal Evolution* (SANE) and *Magnetically Arrested Disks* (MAD) [5, 29, 43, 53]. **We limit our dataset to MAD simulations** as they are currently favored for explaining horizon-scale behavior in both M87* and Sgr A* [54, 56]. In this regime, strong magnetic flux accumulates near the black hole, compressing the accretion disk into a compact, highly dynamic structure. The GRMHD models used in this paper are from the "Illinois v3" library. The fluid simulations are generated using the KHARMA code and imaged with the `ipol` code [42, 45].

Although the same GRMHD framework applies to both M87* and Sgr A*, their observational contexts differ in ways that motivate our approach. M87*, with a mass of $\sim 6.5 \times 10^9$ in solar mass units, evolves on month-long dynamical timescales, making time-averaged simulation images appropriate for comparison with EHT data [15, 18]. In contrast, Sgr A* is $\sim 10^3$ times less massive and varies on minute timescales within a single observing run, where time-averaging washes out physical structure [16, 17]. This rapid variability highlights the need for dynamics-aware methods, such as the framework we develop, that can model and interpret the evolving structure of the accretion flow.