# POLITECNICO DI TORINO

## MASTER's Degree in MATHEMATICAL ENGINEERING


## MASTER's Thesis

# Recovering audio heritage using synchrotron X-rays and machine learning techniques

**Supervisors**

Prof. Marco SCIANNA

Dr. Sebastian GLIGA

**Candidate**

**Lorenzo TERNA**

**MARCH 2026**

**Recovering audio heritage using synchrotron X-rays and machine learning techniques**

**Lorenzo Terna**

**Abstract**

This thesis develops a unified computational pipeline to recover audio from degraded magnetic tapes using the *Play It Again* non-contact readout approach based on X-ray magnetic circular dichroism (XMCD). XMCD acquisitions produce sequences of high-dimensional detector frames in which the magnetization-dependent signal is weak and embedded in strong background and noise. The first contribution is an automated *mask-estimation* method that selects information-bearing pixels on the detector and enables a robust reduction from 2D dichroic frames to a 1D waveform. The mask is learned from a reference dichroic acquisition via unsupervised clustering on physics-inspired pixel features (intensity, local statistics, and gradient-based smoothness), with an optional translation (registration) step to compensate for small footprint drifts. Both binary and confidence-weighted masks are investigated, showing how soft weighting can further suppress background-dominated regions.

To support controlled experimentation and scalable training data generation, the thesis then leverages a stochastic forward emulator of magnetic recording and XMCD readout, modeling key degradation mechanisms such as hysteresis, signal-dependent (heteroscedastic) tape noise, beam blur, and measurement noise. This simulator is integrated into a reproducible HPC workflow to generate large paired datasets, enabling a supervised 1D U-Net denoiser baseline trained with combined time-domain and multi-resolution spectral objectives.

As a future research direction, the restoration problem could be formulated as Bayesian inversion. A differentiable surrogate likelihood is learned to predict the conditional mean and variance of the simulator output, providing stable gradients for likelihood guidance. This is combined with a diffusion prior over clean audio and additional perceptual enhancement terms, yielding guided reverse-diffusion sampling that balances physics consistency with perceptual quality. Overall, the proposed framework bridges XMCD acquisition physics, simulation-driven learning, and probabilistic inference for tape-audio recovery.

ACKNOWLEDGMENTS

*to...*

# Table of Contents

# List of Figures

# List of Tables

# Acronyms

| | |
|---|---|
| PSI | Paul Scherrer Institute. |
| XMCD | X-ray Magnetic Circular Dichroism. |
| SNR | Signal-to-Noise Ratio. |
| ROI | Region of Interest. |
| HPC | High Performance Computing. |
| SLURM | Simple Linux Utility for Resource Management. |
| JIT | Just-In-Time (compilation). |
| FMA | Free Music Archive. |
| GMM | Gaussian Mixture Model. |
| EM | Expectation-Maximization. |
| DBSCAN | Density-Based Spatial Clustering of Applications with Noise. |
| STFT | Short-Time Fourier Transform. |
| MSE | Mean Squared Error. |
| NLL | Negative Log-Likelihood. |
| NRMSE | Normalized Root Mean Square Error. |
| RMSE | Root Mean Square Error. |
| SSIM | Structural Similarity Index Measure. |

| | |
|---|---|
| SI-SDR | Scale-Invariant Signal-to-Distortion Ratio. |
| LSD | Log-Spectral Distance. |
| MFM | Magnetic Force Microscopy. |
| DDPM | Denoising Diffusion Probabilistic Model. |
| GAN | Generative Adversarial Network. |
| AMP | Automatic Mixed Precision. |
| IQR | Interquartile Range. |
| PCM | Pulse-Code Modulation. |
| KL | Kullback-Leibler (divergence). |
| MR-STFT | Multi-Resolution Short-Time Fourier Transform. |

# Part I

# Introduction

# Chapter 1

# Project Context

## 1.1 Play It Again

Magnetic tapes preserve a large fraction of the audio and audiovisual heritage of the last century. They encode information as spatial variations of the remanent magnetization in a thin layer of ferromagnetic particles bound to a polymer substrate. During recording, a write head generates a localized magnetic field that aligns the particle magnetization, creating a pattern of magnetic domains whose polarity and wavelength track the instantaneous signal. For analog audio, the waveform is continuously stored as a modulation of magnetization along the tape; for digital formats, the content is represented by discrete transitions between magnetization states. The recorded pattern is stable in time but can be degraded by aging mechanisms that reduce coercivity, increase noise, or mechanically damage the binder. Many carriers have indeed undergone physical and chemical aging, which makes conventional playback unreliable or destructive: the readout becomes dominated by noise and nonlinear distortion, and in severe cases the tape cannot be safely run through standard heads. The *Play it again* project investigates an alternative, non-contact readout based on X-ray magnetic circular dichroism (XMCD), with the goal of recovering audio, video, and data from degraded tapes [1, 2, 3].

In the experimental workflow, the tape is scanned under controlled X-ray illumination, producing *image-like* detector measurements along the tape trajectory. In XMCD, the absorption of circularly polarized X-rays depends on the relative orientation between the photon helicity and the local magnetization, producing a small but measurable difference in transmitted or emitted intensity. By acquiring images with opposite helicities (or, equivalently, reversing the sample magnetization) and forming their normalized difference, non-magnetic contributions largely cancel while the magnetic contrast changes sign. More precisely, the tape is translated under a focused X-ray beam and the detector records an image sequence along the trajectory, effectively mapping the magnetic microstructure as a function of position. Repeating this procedure under controlled polarization/magnetization conditions yields dichroic images whose contrast is proportional to the component of magnetization probed by the beam, and thus carries the recorded content. These measurements contain a

weak magnetic-contrast component related to the tape magnetization state (hence to the recorded content) up to 4% of total signal, but also strong background structures and detector noise.

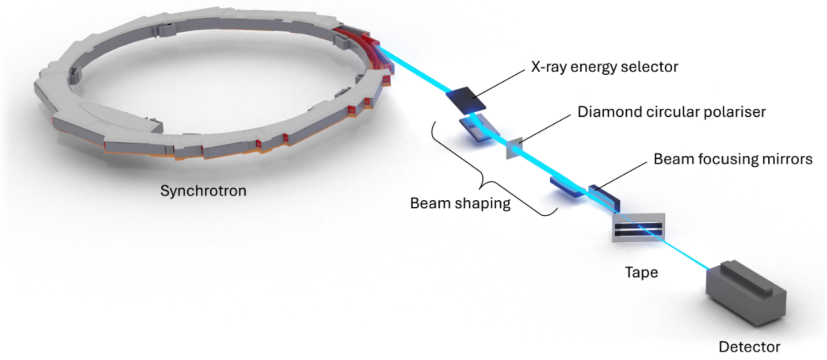


**Figure 1.1:** Schematic of the experimental setup for audio recovery using synchrotron X-ray microspectroscopy.

## 1.2 Motivation

The central computational challenge is to turn a sequence of high-dimensional XMCD frames into a reliable 1D readout, and then to restore the underlying clean audio. A convenient abstraction used throughout this thesis is that each scan index $t \in \{1, \dots, T\}$ provides a 2D frame $I_t \in \mathbb{R}^{H \times W}$ in which only a subset of pixels carries stable dichroic contrast, while the rest is dominated by non-informative background and noise, as anticipated. Consequently, the readout can be modeled as a spatially masked aggregation:

$$y_t \;=\; \langle M_t, I_t \rangle \;=\; \sum_{u=1}^{H} \sum_{v=1}^{W} M_t(u,v)\, I_t(u,v), \tag{1.1}$$

where $M_t$ selects and/or weights informative pixels. This mask viewpoint is not only practical for readout, but also statistically meaningful: it makes explicit how pixel selection affects both the extracted signal and its uncertainty, providing a bridge between experimental design and downstream inference.

Beyond aggregation, the *physics of recording and readout* introduces information loss and stochasticity: hysteresis, granular tape noise, beam blur, and measurement noise. To study these effects and to generate training pairs, has been adopted a stochastic forward model that maps clean audio $x$ to an observed waveform $y$ through latent random variables $\xi$:

$$y = F(x, \xi), \qquad \xi \sim p(\xi). \tag{1.2}$$

The induced restoration problem is an ill-posed statistical inverse problem: recover

$x$ from $y$ in the presence of nonlinearities and uncertainty [4]. This viewpoint motivates combining forward modeling, uncertainty quantification, machine learning and data-driven priors in a single framework.

## 1.3 Proposed objectives

This thesis develops a unified restoration pipeline that connects XMCD readout, physics-based simulation, and Bayesian inference. The objectives are:

1. **XMCD-to-1D signal extraction via mask.** Formalize and construct an automated spatial mask that retains pixels carrying stable dichroic contrast, enabling a robust 1D signal extraction from XMCD frames.

2. **Exploiting stochastic forward modeling for synthetic data and baselines.** Leverage a physics-based simulator of magnetic recording and XMCD readout to generate paired synthetic data $(x, y)$ and establish learning-based denoising baselines.

A further possible extension is a differentiable likelihood surrogate that incorporates mask-induced reliability, and use it to guide a diffusion prior for posterior sampling and perceptual enhancement.

## 1.4 Organization of the thesis

This thesis is structured to progressively develop a complete restoration framework for XMCD-based magnetic tape readout, moving from background concepts to methodological contributions and experimental validation. Chapter 2 introduces the necessary theoretical and technical background, covering magnetic recording principles, XMCD imaging, inverse problems, and modern data-driven restoration methods. Chapter 3 focuses on the first core task, namely the construction of an automated spatial mask for XMCD frames, formalizing the notion of pixel reliability and enabling a robust reduction of high-dimensional images to a one-dimensional signal. In Chapter 4, the attention shifts to forward modeling: a physics-based stochastic simulator of magnetic recording and XMCD readout is exploited, incorporating key effects such as hysteresis, granularity noise, beam blur, and measurement noise, and is used to generate synthetic data for the training of a neural network denoiser. The experimental results corresponding to Tasks 1 and 2 are presented and discussed in Chapters 5 and 6, respectively. Finally, Chapter 7 summarizes the main findings and contributions of the thesis, while Chapter 8 outlines limitations of the current approach and possible directions for future research such as addresses the full inverse problem from a Bayesian perspective, introducing a mask-aware, heteroscedastic likelihood surrogate of the physical acquisition process and combining it with diffusion-based priors to perform guided posterior sampling for perceptual signal enhancement.

# Chapter 2

# Backgrounds

This chapter surveys prior work most closely connected to the thesis, organized around the three pillars of the project: (i) XMCD-based non-contact readout and image-like measurements, (ii) magnetic recording forward modeling with stochastic effects, and (iii) Bayesian inverse problems and modern learned priors for audio restoration, with emphasis on diffusion models and guided posterior sampling.

## 2.1 Non-contact readout and XMCD-based magnetic-contrast imaging

Recovering information from carriers that cannot be safely played back has motivated multiple non-contact or "playback-free" readout strategies, including optical imaging and software-based reconstruction of recorded tracks (e.g., IRENE for grooved media), as well as magnetic-field imaging methods such as magneto-optical indicator films and scanning magnetic microscopies (e.g., MFM or scanning magnetometers) that map the stray-field pattern at high spatial resolution [5, 6, 7, 8]. These approaches, however, typically involve trade-offs between throughput, field-of-view, and robustness to structured backgrounds and ageing artefacts. In contrast, X-ray magnetic circular dichroism (XMCD) is the difference in absorption (or an absorption-related observable) between left- and right-circularly polarized X-rays in magnetic materials, and it changes sign upon reversing either photon helicity or sample magnetization [3, 9]. A common phenomenological model, being $E$ the energy of X-photon (energy of incident radiation), writes an absorption-related signal as

$$\mathcal{S}(E; h, m) \;=\; \mathcal{S}_0(E) + h\, m\, \mathcal{S}_{\mathrm{mag}}(E), \tag{2.1}$$

where $h \in \{+1; -1\}$ is photon helicity and $m \in \{+1; -1\}$ is the magnetization direction (or the controlled field polarity), $\mathcal{S}_0(E)$ is the background term and $\mathcal{S}_{\mathrm{mag}}$ is the XMCD-sensitive term [3, 9]. Equation (2.1) captures the key symmetry: the magnetic term is *odd* in both $h$ and $m$. XMCD contrast arises because the X-ray scattering depends on the relative orientation between the photon angular momentum (helicity) and the sample magnetization. When $m$ or $h$ is reversed, the magnetic

term changes sign, whereas most non-magnetic contributions (intensity variations, detector response, optics) do not. That sign-reversal property makes the reference combination a robust way to isolate the magnetic signal and define "good" pixels.

For each tape index $t$, the experiment yields four detector images

$$I^{(t)}_{h,m} : \Omega \to \mathbb{R}, \qquad (h, m) \in \{+1; -1\}^2.$$

To isolate a magnetic-contrast observable (and to suppress background terms even in helicity and/or magnetization), a double difference is computed

$$D^{(t)}(\mathbf{p}) := \Big(I^{(t)}_{+1,+1}(\mathbf{p}) - I^{(t)}_{-1,+1}(\mathbf{p})\Big) - \Big(I^{(t)}_{+1,-1}(\mathbf{p}) - I^{(t)}_{-1,-1}(\mathbf{p})\Big). \tag{2.2}$$

Under (2.1), Eq. (2.2) enhances the component that is odd in $(h, m)$ and rejects background scattering, providing a *physics-grounded baseline* for subsequent mask construction. The project's experimental workflow produces detector frames (2.1) along the tape scan trajectory, where the desired magnetization-related signal is embedded in strong background contributions and detector noise [1, 2]. Therefore, for the fact that the observations are image-like and must be aggregated into a 1D readout, builing a way to keep only the "good" pixel is necessary. In earlier processing, a mask was obtained via a simple global threshold on a reference dichroic frame, keeping only pixels above (or below) a fixed intensity level and discarding the rest. This proved unreliable because the XMCD contrast is weak and spatially non-uniform, while illumination gradients, thickness/absorption variations, and detector artefacts can dominate the intensity distribution and drift across scans [2, 9]. Consequently, a single threshold is either too conservative (removing magnetically informative pixels and reducing SNR) or too permissive (admitting background-dominated regions and injecting bias), motivating a more robust mask formalization in the next methodology chapter 3.

## 2.2 Magnetic tape recording theory and hysteresis modeling

The forward process from audio drive to recorded magnetization is governed by head-field generation, nonlinear hysteresis, remanence, and saturation. Classical magnetic recording theory provides foundational models for head-field geometry, recording physics, and readout effects [10]. For the head field, Karlqvist-type approximations remain a standard analytical baseline for capturing the spatial decay of the magnetic field produced by the record head [11]. The central nonlinearity is hysteresis, which is the change in the magnetic moment of a material at a given applied magnetic field, based on the hystory of the field experienced by the sample. This is commonly modeled through Preisach operators or related hysteresis frameworks [12]. These references motivate the emulator structure used in Task 2 and explained in the following section.

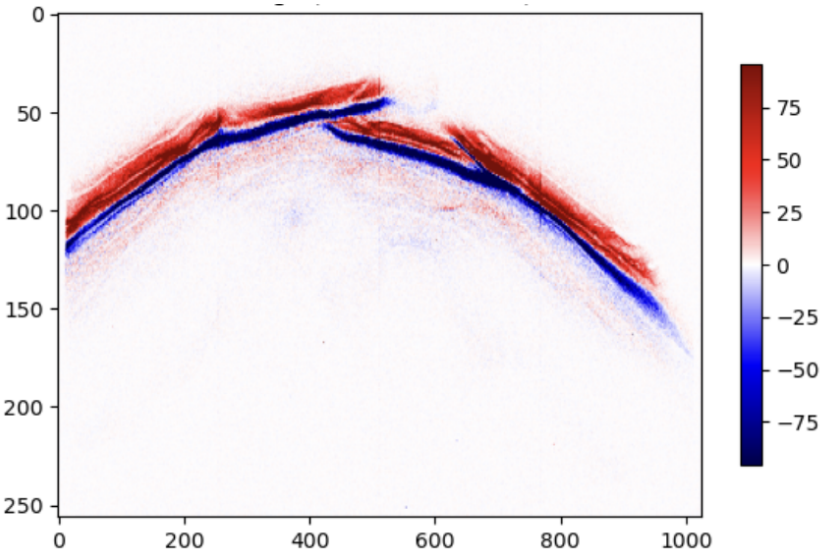


**Figure 2.1:** Combined dichroic image used as reference for mask extraction. The red/blue colormap encodes the sign and magnitude of the dichroic response (red: positive, blue: negative), highlighting the magnetic track as a curved arc across the field of view. The colorbar reports the dichroic intensity in arbitrary units (a.u.).

## 2.3 Emulation Pipeline

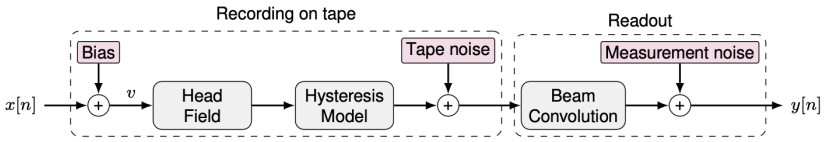


**Figure 2.2:** Emulation pipeline: the input waveform $x[n]$ is joined with the bias, converted to head field at the gap and mapped through a hysteresis model to tape magnetization, including tape noise. It then passes through the beam convolution coming from the XMCD measurement, and finally introduced to measurement noise to yield $y[n]$. Credits for the image: "Developing tools for the recovery of historical audio recordings, João de Azevedo Barbosa"

Let the clean audio clip be $x \in \mathcal{X} := \mathbb{R}^T$ sampled at rate $f_s$. The corresponding observed (degraded) readout is $y \in \mathcal{Y} := \mathbb{R}^T$ after mapping the scan to the same temporal grid used in learning and evaluation. The tape is represented along a 1D spatial coordinate $u$ (aligned with the scan direction) and discretized into $J$ segments, each segment aggregating many magnetic grains. The segment-wise remanent magnetization after recording is denoted by

$$m \in \mathbb{R}^J. \tag{2.3}$$

Assuming constant tape speed $v$, time and space are linked by $u = vt$, allowing spatial operations (e.g., blur) to be expressed on the time axis after sampling.

The simulator is written as a conditional generative model

$$y = F(x, \xi), \qquad \xi \sim p(\xi), \tag{2.4}$$

where $\xi$ collects all stochastic sources (particle packing fluctuations, writing/granularity noise, detector noise, etc.). This viewpoint aligns naturally with likelihood-based inversion [4], useful for further exploiting.

The forward model is expressed as a composition of elementary mappings

$$F(x, \xi) = (f_6 \circ f_5 \circ f_4 \circ f_3 \circ f_2 \circ f_1)(x; \xi), \tag{2.5}$$

where recording physics ($f_1$–$f_3$) is deterministic given parameters, while granularity ($f_4$) and measurement noise ($f_6$) are stochastic. The individual stages are summarized below.

*(1) AC bias addition ($f_1$).* Analog magnetic recording typically uses a high-frequency bias to reduce distortion and linearize the effective transfer characteristic around the working region of the hysteresis loop [10]. This is modeled as

$$x_b(t) = x(t) + A_b \sin(2\pi f_b t), \qquad f_b \gg \text{audio band}, \tag{2.6}$$

so that $f_1(x) = x_b$. *(2) Head field generation ($f_2$).* The biased waveform drives the record head current, and the resulting magnetic field is evaluated along the tape coordinate. A widely used approximation for the head field in the head gap region is given by Karlqvist-type models [10, 11]. Discretizing the tape into $J$ segments yields local field histories $\{H_j(t)\}_{j=1}^{J}$ and a mapping $f_2 :\ x_b \mapsto \{H_j(\cdot)\}$.

*(3) Hysteresis mapping ($f_3$).* Each segment exhibits hysteresis driven by its local field history. Denoting the hysteresis operator by $\mathcal{H}$,

$$M_j^{\text{hist}}(t) = \mathcal{H}[H_j(\cdot)](t), \tag{2.7}$$

where $\mathcal{H}$ can be instantiated through a Preisach model (or related phenomenological models) [10, 12]. The recorded remanence is taken as the end-of-write value $M_j^{\text{hist}} := M_j^{\text{hist}}(t_{\text{end}})$, producing $M^{\text{hist}} \in \mathbb{R}^J$.

*(4) Granular tape noise ($f_4$).* Real tapes are granular: as each segment contains a finite and random number of particles. The realized remanent magnetization therefore fluctuates around the ideal hysteretic prediction. A convenient modeling choice is to treat the particle count and the resulting magnetization as random variables whose variance depends on both packing statistics and the local magnetization level. Concretely, one can sample a segment-wise particle count from packing fluctuations and then model the magnetization variance as proportional to $p_{\text{mag},j}(1 - p_{\text{mag},j})$, where $p_{\text{mag},j}$ is an orientation probability induced by $M_j^{\text{hist}}$. The resulting noisy

remanence is written as

$$m = M^{\text{hist}} + \eta_{\text{tape}}, \qquad \eta_{\text{tape}} \text{ signal-dependent}, \tag{2.8}$$

highlighting that the induced observation model is *heteroscedastic* [13].

*(5) Beam Convolution ($f_5$).* XMCD readout is modeled as a spatial convolution between the remanent magnetization profile and a normalized beam kernel $K$:

$$y_{\text{conv}}(u) = (K * m)(u) = \int K(u-\tau)\, m(\tau)\, d\tau, \qquad where \qquad \int K(u)\, du = 1, \tag{2.9}$$

followed by sampling on the waveform grid via the known scan speed $v$. This step captures the dominant effect of the finite beam footprint, which acts as a low-pass filter on the recovered 1D signal [3].

*(6) Additive measurement noise ($f_6$).* Finally, detector noise and acquisition electronics are modeled as additive noise

$$y = y_{\text{conv}} + w, \qquad w \sim \mathcal{N}(0, \Sigma_{\text{meas}}), \tag{2.10}$$

where $\Sigma_{\text{meas}}$ can be taken as $\sigma^2 I$ or as a diagonal, time-varying covariance if exposure conditions vary along the scan.

The decomposition (2.5) separates (i) smooth, deterministic operators (biasing, head field, blur) from (ii) nonlinear history effects (hysteresis) and (iii) stochastic components (granularity and measurement noise). Crucially, (2.8) implies that uncertainty increases or decreases with local magnetization magnitude, so the conditional distribution $p(y \mid x)$ is not well described by homoscedastic Gaussian noise. This motivates modeling strategies that explicitly account for heteroscedasticity when constructing likelihoods or training conditional models, as discussed in the methodology of Task 3.

## 2.4 Neural network architectures for audio denoising purposes

Neural audio denoisers are commonly trained to recover a clean waveform $x$ from a degraded observation $y$ by learning a parametric mapping $G_\theta : \mathcal{Y} \to \mathcal{X}$ from paired examples. In *speech enhancement* and more general *audio restoration*, two recurring design axes are: (i) the *signal representation* (time-domain waveform vs. time–frequency features) and (ii) the *architecture family* (convolutional encoder–decoders, recurrent models, or hybrids). Time–frequency methods typically estimate a magnitude or complex mask in the STFT domain and reconstruct the waveform via inverse STFT; representative phase-aware examples include DCCRN [14]. In contrast, waveform-to-waveform models operate directly on samples and avoid explicit phase reconstruction, at the cost of modeling long-range temporal structure at high

sampling rates [15, 16].

*Supervised denoising on simulated pairs*

In this thesis, the denoiser used in Task 2 follows the supervised paradigm, but leverages the stochastic acquisition simulator introduced in Section 2.3 as a *paired-data generator*. Clean clips $x \sim p_{\text{data}}$ (in the experiments, drawn from a subset of the Free Music Archive dataset [17]) are degraded through the forward model

$$y = F(x, \xi), \qquad \xi \sim p(\xi), \tag{2.11}$$

where $F$ summarizes the tape/measurement chain (e.g., nonlinear distortion, signal-dependent granularity, blur/low-pass effects, and additive noise; cf. Figure 2.2) and $\xi$ collects its stochastic latent variables. This yields a training set $\mathcal{D} = \{(y_i, x_i)\}_{i=1}^{N}$ with $y_i = F(x_i, \xi_i)$, enabling direct regression training of $G_\theta$ without requiring real ground-truth-aligned XMCD readouts.

The Task 2 objective can be written as empirical risk minimization:

$$\theta^\star \in \arg\min_{\theta} \; \mathbb{E}_{x \sim p_{\text{data}},\, \xi \sim p(\xi)} \Big[ \ell\big(G_\theta(F(x, \xi)),\, x\big) \Big], \tag{2.12}$$

where $\ell$ is a reconstruction loss. The motivation is the *synergy* between simulator and network: the simulator provides abundant, perfectly aligned pairs and controlled variability (through $\xi$ and physics-inspired parameters), while the network amortizes inversion, yielding fast inference applicable at scale.

*A 1D U-Net encoder–decoder for waveform restoration.* A widely adopted backbone for paired restoration is the U-Net encoder–decoder [18], later adapted to 1D signals (Wave-U-Net) for end-to-end audio processing [16]. The key idea is a multi-scale encoder–decoder with *skip connections*: the encoder progressively downsamples the waveform to build large receptive-field features, while the decoder upsamples back to the original resolution and injects fine-scale details by concatenating encoder features at matching scales. This structure is well suited to degradations that mix local artifacts (e.g., transient noise) and broader effects (e.g., blur/low-pass filtering). Figure 2.3 illustrates the resulting "U" shape, with temporal resolution halved at each pooling stage and channel width increased at deeper levels.

In Task 2, $G_\theta$ operates directly in the waveform domain, mapping a mono clip $y \in \mathbb{R}^T$ to an estimate $\hat{x} = G_\theta(y)$. Concretely, the encoder is formed by stacked convolutional blocks—two `Conv1d` layers (kernel size 3, padding 1) with normalization and `LeakyReLU`—followed by `MaxPool1d` downsampling by a factor of 2 at each scale. The decoder mirrors this design, performing upsampling (interpolation followed by `Conv1d`) and concatenating the corresponding encoder activations through skip connections (cf. the lateral links in Fig. 2.3), so that fine time-local structure is preserved while the bottleneck captures long range context. A final $1 \times 1$ convolution maps the last feature map back to a single-channel waveform.

To improve training stability under small batch sizes, affine `GroupNorm` is employed in place of instance normalization. Training is performed with residual learning

enabled: the network predicts a residual $r_\theta(y)$ and outputs $\hat{x} = y + r_\theta(y)$, which typically stabilizes optimization and encourages the model to focus on removing structured corruption rather than re-synthesizing the entire waveform.

*Loss functions: waveform fidelity, spectral structure, and adversarial terms* The simplest choices for $\ell$ in (2.12) are sample-wise losses such as $\ell_2$ (MSE) or $\ell_1$, which encourage pointwise agreement in the time domain. However, perceptual quality often correlates better with matching time–frequency structure. A common compromise is the multi-resolution STFT loss popularized in neural audio synthesis [19]. For several STFT parameterizations $r = 1, \ldots, R$, let $S_r(\cdot)$ denote the complex STFT and $|S_r(\cdot)|$ its magnitude. One combines (i) spectral convergence and (ii) log-magnitude matching:

$$\mathcal{L}_{\mathrm{SC}}^{(r)}(\hat{x}, x) = \frac{\| \, |S_r(\hat{x})| - |S_r(x)| \, \|_F}{\| \, |S_r(x)| \, \|_F}, \tag{2.13}$$

$$\mathcal{L}_{\log}^{(r)}(\hat{x}, x) = \left\| \, \log(|S_r(\hat{x})| + \varepsilon) - \log(|S_r(x)| + \varepsilon) \, \right\|_1, \tag{2.14}$$

and averages over $r$. Here $\| \cdot \|_F$ is Frobenius norm, defined as follow:

given a matrix $A \in \mathbb{R}^{m \times n}$ then $\| A \|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2}$. In practice, Task 2 uses a weighted combination of a time-domain $\ell_1$ term and a multi-resolution STFT term ("aural" loss), enforcing both waveform alignment and spectral plausibility [19]. Notably, losses based purely on magnitude spectra are *not* uniquely identifying: distinct waveforms can share the same $|S_r(\cdot)|$ (e.g. sign inversions or phase variations), which motivates retaining an explicit time-domain fidelity term.

Finally, some approaches incorporate adversarial objectives to reduce over-smoothing and improve realism, as in SEGAN [15] and related conditional GAN formulations [20, 21]. Here, a discriminator learns to distinguish enhanced outputs from real clean audio (in waveform or spectrogram space), while the generator trades off reconstruction and adversarial terms. Although adversarial training can improve perceptual sharpness, it can also introduce instability and hallucinated content; for this reason, it is best viewed as an optional refinement for simulator-trained baselines.

*Why this motivates Task 3: beyond black-box inversion* A conceptual limitation of the Task 2 supervised baseline is that the simulator $F$ is used only *indirectly*, as a mechanism to generate training pairs. The learned inverse $G_\theta$ is not explicitly constrained to respect the pipeline structure, nor to represent physically meaningful intermediate factors (e.g. blur strength or signal-dependent noise levels). Consequently, when the real acquisition deviates from the simulator, performance may degrade because the model has no explicit mechanism to reconcile reconstructions with the known forward process.

This motivates Task 3: rather than relying solely on amortized regression $y \mapsto x$ from simulated pairs, the goal is to perform pipeline-aware inversion, where the forward model and its stochasticity enter *explicitly* into the restoration procedure (e.g. via likelihood-based guidance or other mechanisms that tie reconstructions to

the "soiling" process implemented by $F$).

## 2.5 Summary

This chapter reviewed the foundations most closely connected to non-contact magnetic tape recovery in *Play it again*. First, XMCD provides an element-specific magnetic-contrast mechanism, but the corresponding measurements are *image-like* and often dominated by structured backgrounds and detector noise; as a consequence, early aggregation strategies based on simple global thresholding can be unstable and motivate a more robust masking formalization. Second, classical magnetic recording theory (head-field models and hysteresis frameworks) supports a forward description of how audio is written into remanent magnetization, while realistic tape granularity introduces *signal-dependent* fluctuations that yield heteroscedastic observation uncertainty. These considerations motivate the stochastic emulation pipeline in Section 2.3, which is valuable both for understanding the acquisition process and for generating paired data. Third, modern audio restoration leverages deep neural architectures and learned priors; in particular, the supervised U-Net-style baseline described in Section 2.4 can learn an efficient inverse map from simulated pairs, but it remains a largely *black-box* regressor and does not explicitly enforce consistency with the forward pipeline or expose its uncertainty mechanisms.

This gap motivates Task 3: rather than relying solely on amortized regression, the objective is to incorporate information encoded in the simulation pipeline *directly* into the reconstruction procedure, by defining a mask-aware, heteroscedastic data term (or likelihood surrogate) and combining it with a diffusion prior for guided posterior sampling. Together, these elements bridge XMCD acquisition physics and probabilistic audio restoration, enabling inference that is both physically grounded and uncertainty-aware.

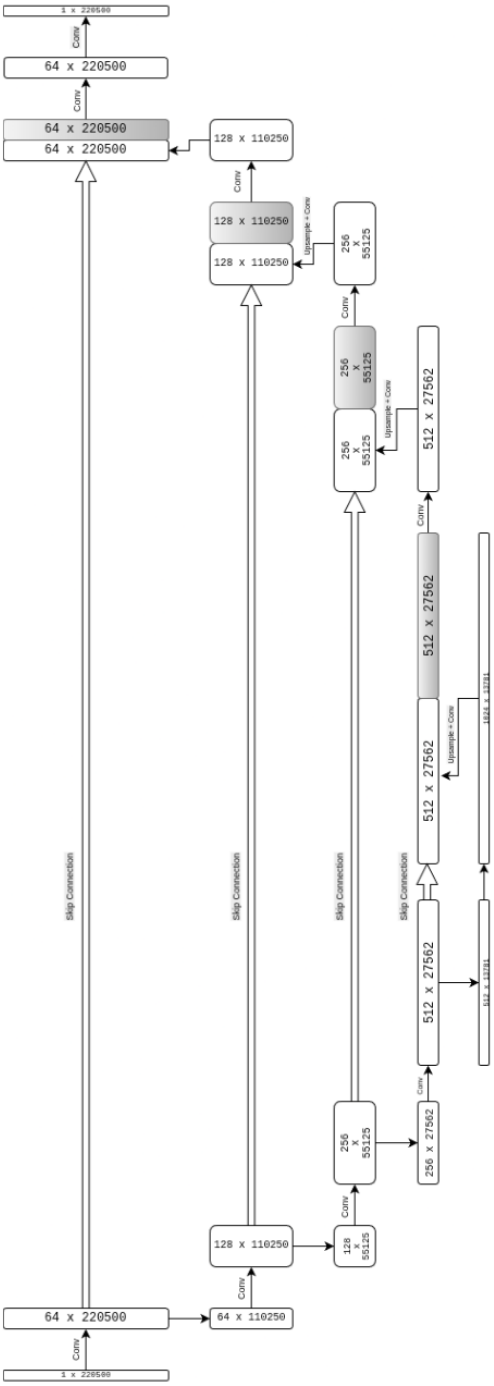


**Figure 2.3:** U-net scheme: A simplified diagram showing the dimensions of the feature maps at each stage of the U-Net architecture Credits for the image: "Deep Learning Models for the Recovery of Magnetic Tape Audio Recordings, Simone Libutti"

# Part II

# Methodology

# Chapter 3

# Task 1: Mask

## 3.1 Project context and problem statement

Section 2.1 motivates the construction of a *mask* that selects information-bearing pixels in an accurate and automatable way.

Let the detector pixel grid be

$$\Omega \;=\; \{1,\dots,H\}\times\{1,\dots,W\}.$$

A mask is a binary function

$$M : \Omega \to \{0;1\}, \tag{3.1}$$

where $M(\mathbf{p}) = 1$ indicates a selected pixel $\mathbf{p} \in \Omega$.

## 3.2 From dichroic frames to a 1D waveform

For each tape position $t$, the dichroic image $D^{(t)}$ is mapped to a scalar sample via a signed aggregation over selected pixels:

$$s_t \;=\; \sum_{\mathbf{p}\in\Omega} M(\mathbf{p})\, D^{(t)}(\mathbf{p}). \tag{3.2}$$

The discrete-time reconstructed signal is the sequence $(s_t)_{t=1}^{T}$.

The central methodological problem is thus: estimate a mask $M$ that retains pixels carrying stable magnetic contrast while rejecting background/noise-dominated pixels, in an automated and reproducible fashion.

## 3.3 Global-mask viewpoint and optional translation operator

*Global mask assumption.* In this thesis, the mask is estimated *globally*: a single mask $M$ is learned once through a reference dichroic image under controlled conditions

where the magnetic contribution is unambiguous (as explained before) and applied to all $t$.

This is justified when the detector footprint of the informative region is stationary across time, or when variations can be captured by a low-dimensional nuisance transformation.

*Time-dependent translation (registration) of the mask.* Although a single global mask is learned from a reference acquisition, in practice the informative footprint can undergo small apparent shifts on the detector due to beam pointing or mechanical drift. To account for this, a per-time translation $\boldsymbol{\delta}_t = (\delta_{t,y}, \delta_{t,x})$ is modeled and define a moving mask through the translation operator

$$M_t(\mathbf{p}) \;:=\; (\mathcal{T}_{\boldsymbol{\delta}_t} M)(\mathbf{p}) \;=\; M(\mathbf{p} - \boldsymbol{\delta}_t), \tag{3.3}$$

so that waveform extraction becomes

$$s_t \;=\; \sum_{\mathbf{p}\in\Omega} M_t(\mathbf{p})\, D^{(t)}(\mathbf{p}), \tag{3.4}$$

that reduces to Eq. (3.2) when $\boldsymbol{\delta}_t = \mathbf{0}$. In this implementation, $\boldsymbol{\delta}_t$ is estimated by registering each acquisition against a chosen reference using phase correlation. Since each NeXus file contains a short stack of detector frames, first the stack is averaged to obtain a single representative image per acquisition. To reduce sensitivity to file-to-file gain and offset changes, images are robustly normalized by subtracting the median and dividing by a robust scale (IQR, with a percentile fallback when needed). The phase-correlation peak yields the shift estimate, refined at subpixel resolution via local upsampling; for diagnostics the estimated transformation is applied using bilinear interpolation (equivalently, one may translate the mask under the same interpolation convention). It's quantified whether the moving-mask correction is necessary by comparing overlap metrics before and after alignment, using correlation, SSIM, and a pooled-scale normalized RMSE

$$\mathrm{NRMSE}_{\mathrm{pooled}}(A, B) \;=\; \frac{\sqrt{\mathbb{E}[(A-B)^2]}}{\sqrt{\mathrm{Var}(A) + \mathrm{Var}(B)}}, \tag{3.5}$$

computed either globally or within a footprint ROI when available. To disentangle geometric drift from purely photometric differences, additionally an affine intensity model $\bar{D}^{(t)} \approx a\,\bar{D}^{(\mathrm{ref})} + b$ is fitted and checked how much it reduces the residual error; simple second-moment descriptors (centroid and spread) is also monitored to detect non-translational footprint changes. In the remainder of this thesis, the moving mask is enabled only when $\hat{\boldsymbol{\delta}}_t$ is non-negligible and alignment yields a consistent improvement in these diagnostics; otherwise the global mask is applied unchanged.

## 3.4 Feature embedding on the dichroic image

A key hypothesis motivating unsupervised learning here is that signal-carrying pixels exhibit a *structured spatial distribution* induced by the underlying physics and imaging geometry. Therefore, clustering should not rely on raw intensity alone, but should incorporate local spatial statistics.

*Reference image for feature computation.* Concretely, a reference magnetized frame $D^{(\mathrm{ref})}$ used to build the global mask is chosen.

Two common choices are:

- A representative dichroic frame: $D^{(\mathrm{ref})} := D^{(t_0)}$ for some index $t_0$.
- A robust aggregate (stability-enhancing): e.g. pixelwise median

$$D^{(\mathrm{ref})}(\mathbf{p}) := \mathrm{median}_{t\in\mathcal{I}}\, D^{(t)}(\mathbf{p})$$

  over a subset $\mathcal{I} \subset \{1, \ldots, T\}$.

The methodology below is agnostic to the choice.

Let $X(\mathbf{p}) := D^{(\mathrm{ref})}(\mathbf{p})$.

*Five-dimensional pixel feature map.* For each pixel $\mathbf{p} \in \Omega$, define the feature vector $\mathbf{f}(\mathbf{p}) \in \mathbb{R}^5$:

$$\mathbf{f}(\mathbf{p}) = \big[f_1(\mathbf{p}),\ f_2(\mathbf{p}),\ f_3(\mathbf{p}),\ f_4(\mathbf{p}),\ f_5(\mathbf{p})\big]^\top, \tag{3.6}$$

where:

$$f_1(\mathbf{p}) = X(\mathbf{p}), \tag{3.7}$$

$$f_2(\mathbf{p}) = |X(\mathbf{p})|, \tag{3.8}$$

$$f_3(\mathbf{p}) = \mu_r(\mathbf{p}) = \frac{1}{|N_r(\mathbf{p})|} \sum_{\mathbf{q}\in N_r(\mathbf{p})} X(\mathbf{q}), \tag{3.9}$$

$$f_4(\mathbf{p}) = \sigma_r(\mathbf{p}) = \sqrt{\frac{1}{|N_r(\mathbf{p})|-1} \sum_{\mathbf{q}\in N_r(\mathbf{p})} \Big(X(\mathbf{q}) - \mu_r(\mathbf{p})\Big)^2}, \tag{3.10}$$

$$f_5(\mathbf{p}) = \|\nabla X(\mathbf{p})\|_2. \tag{3.11}$$

Here $N_r(\mathbf{p})$ denotes a neighborhood window of radius (or half-width) $r$ around $\mathbf{p}$, and the gradient $\nabla X$ is computed via a discrete derivative operator [22, 23].

*Interpretation.* The pair $(f_1, f_2)$ captures signed contrast and magnitude (important because XMCD contrast can change sign depending on local magnetization). The local statistics $(f_3, f_4)$ encode spatial continuity and suppress isolated noise. The gradient magnitude $f_5$ penalizes high-frequency structures and helps discriminate smooth footprints from spurious speckle-like artifacts [22].

*Feature scaling.* Because the coordinates in (3.6) have heterogeneous scales, features

are standardized:

$$\tilde{f}_k(\mathbf{p}) = \frac{f_k(\mathbf{p}) - \mathbb{E}_{\mathbf{p}}[f_k(\mathbf{p})]}{\sqrt{\mathrm{Var}_{\mathbf{p}}[f_k(\mathbf{p})]} + \varepsilon}, \qquad k = 1, \ldots, 5, \tag{3.12}$$

yielding normalized vectors $\tilde{\mathbf{f}}(\mathbf{p})$ for clustering.

## 3.5 Unsupervised clustering methods for mask discovery

Let $\mathcal{F} = \{\tilde{\mathbf{f}}(\mathbf{p})\}_{\mathbf{p}\in\Omega}$ be the standardized feature cloud. Three complementary clustering paradigms are considered.

*k-means.* $k$-means partition $\mathcal{F}$ into $K$ clusters by minimizing within-cluster squared Euclidean distances [24, 25]:

$$\min_{\{\mathcal{C}_k\}_{k=1}^{K}} \sum_{k=1}^{K} \sum_{\mathbf{p}\in\mathcal{C}_k} \|\tilde{\mathbf{f}}(\mathbf{p}) - \boldsymbol{\mu}_k\|_2^2, \tag{3.13}$$

with centroids $\boldsymbol{\mu}_k$. After fitting, a signal cluster set $\mathcal{K}_{\mathrm{sig}}$ is chosen, and

$$M(\mathbf{p}) = \mathbb{I}\{\mathrm{cluster}(\mathbf{p}) \in \mathcal{K}_{\mathrm{sig}}\}.$$

*Gaussian mixture models (GMM) with EM.* A GMM models the feature distribution as a mixture of Gaussians:

$$p(\tilde{\mathbf{f}}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\tilde{\mathbf{f}}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad \sum_k \pi_k = 1,\ \pi_k \geq 0. \tag{3.14}$$

Parameters are estimated by the expectation–maximization algorithm [26]. The posterior responsibility

$$\gamma_k(\mathbf{p}) = p(k \mid \tilde{\mathbf{f}}(\mathbf{p}))$$

enables either a hard mask (MAP assignment) or a soft mask threshold:

$$M(\mathbf{p}) = \mathbb{I}\Big\{ \sum_{k\in\mathcal{K}_{\mathrm{sig}}} \gamma_k(\mathbf{p}) \geq \tau \Big\}. \tag{3.15}$$

*DBSCAN (density-based clustering with noise labeling).* DBSCAN identifies clusters as high-density regions in feature space and labels low-density points as noise [27]. It depends on $\varepsilon$ (neighborhood radius) and minPts. A point is a core point if at least minPts neighbors lie within $\varepsilon$. Pixels labeled as DBSCAN noise are naturally rejected (set to 0 in $M$), which matches the masking objective.

## 3.6 Mask establishment

Once clustering has been performed on the reference dichroic image $D^{(\mathrm{ref})}$, each pixel $\mathbf{p} \in \Omega$ is assigned a cluster label $\ell(\mathbf{p}) \in \{1, \ldots, K\}$. The clustering step itself is

agnostic to the physical meaning of the XMCD contrast; it only groups pixels with similar feature statistics.

To establish the final masks, clusters are ranked according to their mean dichroic intensity on the reference image. For each cluster $k$, is computed

$$\mu_k = \frac{1}{|\Omega_k|} \sum_{\mathbf{p} \in \Omega_k} D^{(\mathrm{ref})}(\mathbf{p}), \tag{3.16}$$

where $\Omega_k = \{\mathbf{p} \in \Omega : \ell(\mathbf{p}) = k\}$.

The cluster with the largest mean value $\mu_k$ is selected as the *positive mask*, while the cluster with the smallest mean value is selected as the *negative mask*. All remaining clusters, whose mean dichroic contrast lies between these two extremes, are considered background or noise and are discarded.

This ranking-based selection exploits the known sign structure of the XMCD signal in the reference acquisition and yields a simple, fully automatic separation between positive, negative, and non-informative pixels.

## 3.7 Weighted Mask

The binary mask treats all selected pixels equally, which is suboptimal when the footprint contains regions with different signal-to-noise ratios. The weighted mask replaces the discrete values $\{-1; 0; +1\}$ with a continuous weight

$$W : \Omega \to [-1, +1],$$

so that pixels that are *more likely* to carry magnetic contrast contribute more to the waveform reconstruction.

*Soft membership (confidence) from clustering.* In the GMM case, each pixel has responsibilities $\gamma_k(\mathbf{p}) = p(k \mid \tilde{\mathbf{f}}(\mathbf{p}))$ that naturally come from the posterior probabilities of the GMM model.

For hard-label clusterers (as $k$-means), an analogous confidence can be obtained from distances-to-centroids (for $k$-means) or from a density score.

*From $p_{pos}$ and $p_{neg}$ to a signed weighted mask.* After fitting either *K-means* (with distance–softmax posteriors) or a *Gaussian Mixture Model* (with true posterior responsibilities), each valid pixel $i \in \{1; \ldots; N\}$ is associated with two scalar scores:

$$\gamma_{\mathrm{pos}}(p) \in [0, 1], \qquad \gamma_{\mathrm{neg}}(p) \in [0, 1],$$

where $\gamma_{\mathrm{pos}}(p)$ is the probability (or soft assignment) that pixel $p$ belongs to the cluster whose mean dichroic signal $D$ is maximal, and $\gamma_{\mathrm{neg}}(p)$ analogously corresponds to the cluster whose mean $D$ is minimal.

*Signed membership (polarity estimate).* To obtain a signed indicator of polarity, the difference

$$s(p) \;=\; \gamma_{\mathrm{pos}}(p) - \gamma_{\mathrm{neg}}(p) \;\in\; [-1, 1]. \tag{3.17}$$

is taken. Pixels with $s(p) \approx +1$ are strongly associated with the positive-contrast extreme, while $s(p) \approx -1$ indicates association with the negative-contrast extreme. Values near 0 correspond to ambiguous pixels or background.

*Extreme-class mass (background suppression).* Since many pixels may be explained by *intermediate* clusters (background / texture), the total mass assigned to the two extremes is computed:

$$m(p) \;=\; \gamma_{\mathrm{pos}}(p) + \gamma_{\mathrm{neg}}(p) \;\in\; [0,1]. \tag{3.18}$$

If a pixel belongs to neither extreme, then both posteriors are small and $m(p)$ is close to 0, automatically suppressing uninformative regions.

*Signed confidence score.* Polarity and extreme-class mass are combined via

$$c(p) \;=\; s(p)\, m(p) \;=\; \big(\gamma_{\mathrm{pos}}(p) - \gamma_{\mathrm{neg}}(p)\big)\big(\gamma_{\mathrm{pos}}(p) + \gamma_{\mathrm{neg}}(p)\big) \;=\; \big(\gamma_{\mathrm{pos}}(p)\big)^2 - \big(\gamma_{\mathrm{neg}}(p)\big)^2, \tag{3.19}$$

which remains in $[-1,1]$ and emphasizes pixels that are both (i) strongly polarized toward one extreme and (ii) confidently assigned to the set of extreme clusters, while attenuating pixels explained by intermediate/background components.

*Amplitude gating with robust normalization.* To further down-weight low-contrast pixels (likely dominated by noise), an amplitude gate $a(p) \in [0,1]$ is introduced based on the absolute dichroic value $|D(p)|$:

$$a(p) \;=\; \mathrm{clip}\!\left(\frac{|D(p)|}{h},\, 0,\, 1\right)^{\gamma}, \qquad h \;=\; \mathrm{perc}_q(|D|), \tag{3.20}$$

where $\mathrm{perc}_q(\cdot)$ denotes a high percentile ($q = 99.5$) used as a robust scale estimate, $\gamma$ is an optional nonlinearity, and an optional floor percentile may set very small $|D(p)|$ to zero before normalization.

*Final weights, spatial mapping, and smoothing.* The final per-pixel weight is computed as

$$w(p) \;=\; c(p)\, a(p) \;\in\; [-1,1]. \tag{3.21}$$

Weights are written back to the image grid by assigning $w(p)$ to valid pixels and 0 to invalid/saturated pixels. Finally, an optional Gaussian smoothing enforces spatial coherence:

$$W \;=\; \mathcal{G}_\sigma * W_{\mathrm{raw}},$$

followed by clipping to $[-1,1]$. The resulting mask $W$ encodes both XMCD polarity (sign) and a continuous confidence score (magnitude), enabling a soft selection of magnetically informative pixels instead of a hard binary mask.

## 3.8 Summary

This chapter formalized a physics-grounded, unsupervised approach to detector pixel selection for XMCD tape readout. A double-difference dichroic image (2.2) isolates a helicity- and magnetization-odd contrast component [3, 9]. Features computed directly on the dichroic reference image encode intensity, magnitude, spatial coherence, and smoothness. Clustering methods ($k$-means, GMM/EM, DBSCAN) identify a candidate signal footprint, which is refined by spatial regularization. A single global mask is learned and optionally translated by a time-dependent operator to account for drift, enabling robust waveform extraction by signed aggregation.

# Chapter 4

# Task 2: Emulation pipeline and preliminary denoiser merging as denoiser baseline

## 4.1 Dataset

To evaluate the denoising baseline on a stylistically coherent domain while keeping iteration times manageable during development, two complementary corpora derived from the Free Music Archive (FMA) have been used. FMA is a large-scale, openly licensed dataset designed for Music Information Retrieval, providing audio together with rich track-level metadata and a hierarchical genre taxonomy. The first corpus is FMA-small, which provides a lightweight subset with the same overall structure and metadata format, allowing access to more generalized data. The other dataset used in Task 2 is built from the *audio tracks of FMA-large*, filtered to retain only Classical material using the official metadata archive (`fma-metadata.zip`): the genre identifier(s) corresponding to the label "Classical" are retrieved from `genres.csv`, then candidate tracks are selected from `tracks.csv` by checking whether those identifiers appear in the per-track genre list (e.g., `track.genres_all`), which is robust to the hierarchical taxonomy and retains tracks tagged as Classical even when additional subgenre assignments are present. From the resulting pool, a random sample of 1,602 tracks is drawn and split into train/validation/test (80/10/10) *at the track level* using a fixed random seed, which prevents leakage from multiple clips of the same track across splits; for exact reproducibility, explicit track-id lists per split are stored and reused in all experiments. The subset is stored on disk following the original FMA layout, placing each MP3 under a three-digit directory computed from its track identifier (e.g., `000/`, `001/`, ...) to keep directory sizes manageable and to remain compatible with FMA scripts; the final structure is therefore `classical-1602/{train,val,test}/{000,001,...}/*.mp3`, with mutually exclusive track lists and an explicit mapping file supporting replication. Since the raw corpus is MP3 while both emulator and denoiser operate on waveform tensors, audio is decoded using `torchaudio` and each track is converted to mono by channel

averaging; generated clips are saved as 32-bit PCM `.wav` to avoid quantization artifacts and ensure consistent I/O across machines. All tracks are treated as 30-second excerpts and a strict target duration of 30 seconds is enforced: tracks longer than 30 seconds are truncated, while slightly shorter tracks are zero-padded up to the target length within a tolerance threshold (0.5 s); tracks shorter than this tolerance are discarded to prevent excessive padding from dominating the signal. Each track is split into fixed-length clips of duration 5 seconds, and for each clean clip $x$ the pipeline can produce a degraded observation $y = F(x, \xi)$ via the physical emulator. In the current implementation, paired clips are generated at a fixed working sample rate of 48 kHz to match emulator assumptions and internal discretization, while the denoiser training pipeline uses a 44.1 kHz training rate; therefore the paired loader optionally resamples *both* $x$ and $y$ on-the-fly to the training rate, ensuring consistent tensor shapes in every batch (this design choice is revisited in Chapter 6 through dedicated ablations). Reproducibility is enforced at two levels: the track-level split is deterministic due to the fixed seed and the stored track-id lists, and the emulator stochasticity is made deterministic *per clip* by deriving a unique seed from a stable identifier (relative path and clip index), so that datasets can be regenerated bit-for-bit regardless of parallel execution order on the cluster. Finally, each generated dataset folder stores a `params.json` snapshot with preprocessing and emulation parameters (sample rate, clip length, bias configuration, filter cutoffs, and physical knobs such as beam size, step size, and segment length), making synthetic datasets self-describing and enabling lightweight provenance tracking.

## 4.2 Cluster Merlin7

The main practical challenge of Task 2 is computational: generating a paired dataset requires running the full physics-based forward simulator on a large number of short audio clips, and the hysteresis stage dominates runtime. For this reason, dataset generation and baseline training were executed on the Merlin7 HPC infrastructure using distinct CPU and GPU job profiles, structured as a two-stage workflow that separates (i) *paired-data generation* on CPU nodes from (ii) *denoiser training* on GPU nodes, so that each stage can be validated independently and matched to its resource requirements. Paired-data generation is executed as a SLURM array job on the `general` partition: the script enumerates audio files under each split directory, partitions the list into `K` disjoint fractions, and assigns one fraction to each array task via `--which-fraction` (in the configuration considered `K=4` and array `1-4`), ensuring non-overlapping writes in the output directory tree. To use allocated CPU cores efficiently and avoid oversubscription, threading environment variables (`OMP_NUM_THREADS`, `MKL_NUM_THREADS`, `OPENBLAS_NUM_THREADS`, `NUMEXPR_NUM_THREADS`) are set to match `--cpus-per-task`, [28].; since parts of the simulator are accelerated with Numba, a just-in-time (JIT) compiler for Python that translates a subset of Python/NumPy code into optimized machine code at runtime, the hottest numerical kernels (e.g. tight loops over arrays) can run at near-native

**Table 4.1:** HPC job configurations used in Task 2. CPU array jobs generate paired clips; GPU jobs train and validate the baseline denoiser.

| Stage | Cluster | Partition | Resources | Wall-time |
|---|---|---|---|---|
| Clip generation (paired data) | `merlin7` | `general` | Array 1–4; 1 task; 16 CPU; 64 GB RAM | 4 days |
| Training (full run) | `gmerlin7` | `a100-daily` | 1×GPU (A100); 8 CPU; 32 GB RAM | 24 hours |
| Training (smoke test) | `gmerlin7` | `a100-general` | 1×GPU (A100); 4 CPU; 16 GB RAM | 10 minutes |

speed without rewriting them in C/C++ [29, 30]. To make performance predictable on HPC nodes, the compilation cache is placed on node-local scratch (`SLURM_TMPDIR`) via `NUMBA_CACHE_DIR`, so that compiled artifacts are written to a fast local filesystem and reused across repeated calls within the same job [31]. Also a short warm-up call has been included so that JIT compilation happens early and runtime remains stable across clips. Training runs are then executed on GPU nodes (1×A100), using a longer wall-time allocation for full runs and a short smoke-test profile for end-to-end validation after code changes, with structured logs and checkpoints written to dedicated folders controlled through environment variables (e.g., `CHECKPOINT_FOLDER` and `DATA_DIRECTORY`).

This cluster setup enabled repeated integration iterations while preserving a reproducible execution trail (array logs, configuration snapshots, deterministic seeding, and stable dataset naming). The resulting paired datasets and trained checkpoints provide the supervised denoiser baseline for Task 2, while the runtime bottleneck of the forward model motivates the hysteresis-reduction strategy described next.

## 4.3 Merging Implementation

Task 2 combines two pre-existing codebases: (i) a physics-based forward simulator (emulation pipeline) that maps clean audio to an XMCD-like degraded readout, and (ii) a supervised audio denoiser baseline (Wave-U-Net style) that expects large collections of paired training examples. While conceptually complementary, the two implementations differed in their I/O conventions (file layout, clip extraction strategy, sampling rate), stochasticity handling, and computational assumptions. The main methodological contribution of this task is therefore an integration layer that makes the physical simulator usable as a drop-in data generator for the denoiser training pipeline, while preserving reproducibility and enabling scalable execution on a cluster.

Concretely, the merged implementation introduces: (a) a deterministic seeding strategy that makes generated datasets exactly reproducible, (b) a controllable reduction of the hysteresis cost through coarser spatial discretization, and (c) a robust paired-data loader compatible with the hierarchical layout of FMA-derived datasets.

*Hysteresis reduction* The hysteresis block is the dominant computational bottleneck

of the emulation pipeline. In the merged implementation, its cost is reduced through two complementary approximations: (i) reducing the number of Preisach hysterons (*coarse Preisach model*).

The original Preisach fit stores a large set of hysterons, each parameterized by its switching thresholds $(\alpha_h, \beta_h)$ and an associated signed weight $w_h$. To obtain a computationally cheaper model while preserving the global nonlinear transfer characteristic, a coarse representation is built by binning the $(\alpha, \beta)$ plane into an $N \times N$ grid (with $N = 250$ in the main configuration). For each bin $b$, the following quantities are computed:

$$w_b = \sum_{h \in b} w_h, \tag{4.1}$$

$$\alpha_b = \frac{\sum_{h \in b} \alpha_h |w_h|}{\sum_{h \in b} |w_h|}, \qquad \beta_b = \frac{\sum_{h \in b} \beta_h |w_h|}{\sum_{h \in b} |w_h|}. \tag{4.2}$$

Using $|w_h|$ in the averages prevents cancellations between positive and negative weights when selecting the representative $(\alpha_b, \beta_b)$, while $w_b$ retains the signed contribution of the bin to the final magnetization. Bins with $w_b \neq 0$ are retained, yielding a reduced hysteron set $\{(\alpha_b, \beta_b, w_b)\}_b$ saved as a compact `.npz` resource. The reversible susceptibility term ($\chi_{\text{rev}}$) is carried over unchanged and stored alongside the coarse arrays.

The hysteresis simulator evaluates the remanent magnetization on a 1D tape patch whose length is proportional to the clip duration $T$ and tape speed $v$. The number of simulated spatial points scales approximately as $N_x \approx vT/\Delta\ell$, where $\Delta\ell$ is the tape segment length. Increasing $\Delta\ell$ reduces $N_x$ linearly and therefore lowers the total cost of the Preisach evaluation over space. In the merged codebase, $\Delta\ell$ is exposed as a single knob (`segment-length-um`) so that the fidelity–runtime trade-off can be explicitly controlled and systematically evaluated. One aims to use the greatest value possible while keeping physycal realiablity; that is why in the main configuration the `segment-length-um` is chosen equal to `head-gap` ($= 4\mu m$).

Overall, coarse Preisach reduction lowers the cost per hysteresis evaluation, while increasing $\Delta\ell$ reduces the number of evaluations per clip. Together, these changes enable dataset-scale generation on Merlin7 while maintaining the key nonlinear and stochastic characteristics needed for training a meaningful denoiser baseline.

*Unified clip-wise data generation* The original emulator was refactored from a single-file workflow into a callable, clip-wise transformation that can be used inside a dataset generation script. Each audio track is loaded from disk, converted to mono, and resampled to a fixed working sample rate (48 kHz in the generation pipeline). Tracks are then brought to a fixed duration (30 s), by truncation or zero-padding within a tolerance threshold, and split into fixed-length clips.

If the `--apply-transforms` flag is enabled, each clean clip $x$ is passed through the physical transformation $y = F(x, \xi)$ implemented by the merged emulator. The transformation includes bias handling (off/fixed/"3std" modes), and the stochastic noise sources included in the forward model. To reduce overhead, the emulator object

is instantiated once and reused across clips (cached in-memory), while Numba kernels are compiled during an explicit warm-up step.

The script mirrors the directory structure of the clean dataset when writing outputs. This produces a paired dataset where the degraded input and clean target share the same relative path under their respective roots, enabling robust 1:1 pairing without relying on file name heuristics.

*Deterministic stochasticity and reproducibility* Because the forward model is stochastic, controlling randomness is essential to obtain reproducible datasets and to enable consistent training comparisons. Determinism is enforced at the *clip level* by deriving a unique seed for each clip from a stable identifier. Concretely, for each track a relative path key is computed and combined with the clip index to form a string identifier. A 32-bit CRC hash of this identifier is then added to a global seed:

$$s_{\text{clip}} = s_0 + \text{CRC32}(\text{relpath} \,||\, \text{clip\_index}), \tag{4.3}$$

and $s_{\text{clip}}$ is passed to the emulator. This ensures that (i) the same dataset can be regenerated bit-for-bit, and (ii) parallel generation via SLURM arrays yields identical results to a single-process run, since each clip's randomness is independent of execution order.

To make dataset instances self-describing, a `params.json` file is stored next to the generated clips, capturing all relevant preprocessing and emulator parameters (sample rate, clip duration, beam size, step size, segment length, bias settings, and filter settings).

*Denoiser-side integration: paired loader and training entrypoints* On the denoiser side, the key requirement is a robust pairing mechanism between degraded inputs and clean targets. A paired dataset loader is implemented that (i) resolves split directories in a case-insensitive way (`train/Train/TRAIN`, etc.), and (ii) constructs 1:1 pairs by mapping each noisy clip to its clean counterpart through the relative path within the split directory. This approach remains valid even when the dataset is stored using nested subfolders (as in FMA-like layouts), and it naturally supports large datasets without building fragile global file lists.

The loader optionally resamples waveforms on-the-fly to the denoiser training sample rate (44.1 kHz in the baseline configuration), converts to mono if needed, and applies a safety trim to enforce identical lengths for $(y, x)$ within a pair. For convenience, the training script accepts either (a) a full path to the noisy dataset root, or (b) a short dataset identifier that is expanded to the corresponding folder name; the clean target root is then inferred automatically from the clip duration.

Training is executed as a single-GPU job on the A100 partition with standard checkpointing and logging. In addition to full training runs (150 epochs), a short "smoke test" configuration (1 epoch, small batch size, short wall-time) is maintained to validate end-to-end correctness after integration changes (data loading, forward pass, loss computation, checkpoint writing) before launching longer runs.

## 4.4 Denoiser baseline

As a supervised baseline for Task 2, the waveform denoiser implementation developed in prior work is adopted and adapted to the paired datasets produced by the merged emulation pipeline. The baseline is a 1D U-Net operating directly on raw audio waveforms: it maps a noisy mono waveform to a predicted correction signal and reconstructs the clean waveform via residual learning. Concretely, the network follows the design with skip connections presented in Ch. 2. Training uses a perceptually motivated combined loss (`AuralLoss`) mixing waveform L1 with a multi-resolution STFT loss,

$$\mathcal{L}(\hat{x}, x) = (1 - \alpha)\, \|\hat{x} - x\|_1 + \alpha\, \mathcal{L}_{\text{MR-STFT}}(\hat{x}, x), \tag{4.4}$$

with $\alpha = 0.3$ in the baseline configuration; the MR-STFT term averages spectral convergence and log-magnitude losses over multiple FFT/hop/window settings and is evaluated in FP32 for numerical stability under mixed precision. For inference on longer recordings, overlap-add is used: the input is split into chunks, each chunk is denoised independently.

## 4.5 Training protocol

The baseline is trained with a supervised objective on paired clips $(y, x)$, where $x$ is the clean target and $y$ is its degraded counterpart produced by the physical emulator, minimizing the chosen loss $\mathcal{L}(\hat{x}, x)$ over the training split with $\hat{x} = G_\theta(y)$ predicted by the 1D U-Net. During loading, waveforms are converted to mono if needed and safety-trimmed to enforce identical lengths for $(y, x)$, and when generating at 48 kHz but training at 44.1 kHz the loader deterministically resamples *both* members of each pair using the same operator to preserve supervised consistency. Optimization uses Adam with mild weight decay, a learning rate of $3 \times 10^{-5}$ and a cosine scheduler, with full runs up to 150 epochs (batch size 16) on a single A100 GPU, alongside a short smoke-test configuration (1 epoch, small batch size, short wall-time) to validate end-to-end correctness after integration changes. Training uses automatic mixed precision (AMP) with gradient scaling for throughput, global gradient clipping (norm capped at 1.0) for stability with STFT-based losses, and `torch.compile` when supported to reduce Python overhead. At the end of each epoch the average validation loss on the validation split is computed, losses, learning rate, and waveform statistics (RMS and peak) are logged to TensorBoard, and checkpoints containing model/optimizer/AMP scaler/scheduler states and tracked metrics are saved; if a checkpoint for the requested configuration exists, training resumes automatically, enabling robust long runs under time-limited allocations.

## 4.6 Summary

Task 2 establishes a supervised denoising baseline by combining a physics-based emulation pipeline with a waveform denoiser trained on synthetically generated paired data. The main outcome is an end-to-end, reproducible workflow that produces aligned noisy/clean clip pairs at scale (CPU stage) and trains a 1D U-Net denoiser on these pairs (GPU stage).

While this approach enables controlled experiments, it may still suffer from a *domain gap*: the denoiser is trained on samples from the emulator distribution, which may not perfectly match real XMCD degradations. Moreover, practical engineering choices adopted for feasibility, such as hysteresis reduction and generating data at 48 kHz while training at 44.1 kHz, an slightly affect the synthetic distribution and therefore the learned restoration mapping. These limitations motivate Task 3, whose goal is to incorporate the forward-model information more directly into the restoration procedure, rather than using it only implicitly through supervised synthetic training pairs.

# Part III

# Experiments and Results

# Chapter 5

# Task 1: Experiments and Results

This chapter reports the empirical validation of the mask-estimation pipeline introduced in Chapter 3. The goal is twofold: (i) verify that unsupervised clustering on a reference dichroic frame produces a spatially coherent footprint consistent with the expected XMCD contrast distribution; and (ii) quantify how the resulting (binary or weighted) mask impacts the reconstructed 1D waveform via Eq. (3.2) (and, when enabled, Eq. (3.4)).

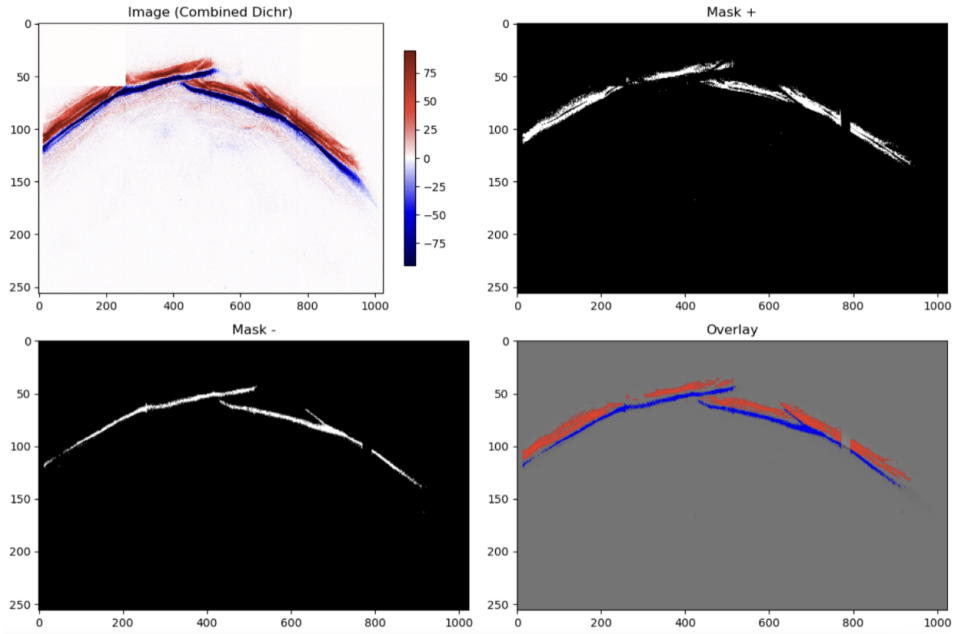


**Figure 5.1:** Reference dichroic frame and corresponding mask overlays obtained with the tested clustering K-means back-end. The comparison highlights the spatial location, compactness, and fragmentation of the estimated informative footprint for the used method.

## 5.1 Experimental protocol

*Data and reference acquisitions.* All experiments use NeXus acquisitions in which each file contains a short stack of detector frames. For each acquisition, the stack is

averaged to obtain a single representative dichroic image. Saturated/invalid pixels are excluded from all feature computations and clustering.

A dedicated *reference* acquisition (magnetized, high-contrast, controlled conditions) is used to build the global mask, yielding $D^{(\mathrm{ref})}$ (Section 3.4), per each experiment. When multiple candidates are available, $D^{(\mathrm{ref})}$ is chosen to be representative of the nominal footprint and to maximize contrast-to-noise.

Unless otherwise stated, robust normalization is applied by subtracting the pixelwise median and dividing by a robust interquartile range scale (IQR), with a percentile fallback when IQR is numerically small (Chapter 3).

*Feature and clustering configurations.* Pixel features are computed from the reference image $X(\mathbf{p}) := D^{(\mathrm{ref})}(\mathbf{p})$ using the five-dimensional embedding in Eq. (3.6). The neighborhood radius $r$ in Eqs. (3.9)–(3.10) is set to $r = 7$ *pixels*; the gradient magnitude uses a discrete derivative operator (Sobel) as in Eq. (3.11). Features are standardized as in Section 3.4.

Three clustering back-ends are evaluated (Section 3.5):

- **$k$-means:** number of clusters $K = 3$; $n_{\mathrm{init}} = 10$; fixed random seed.
- **GMM/EM:** $K = 3$; covariance type *full*.
- **DBSCAN:** $\varepsilon = 0.75$; minPts $= 25$.

*Mask variants.* The following variants are compared:

1. **Binary global mask** $M$ from cluster ranking (Section 3.5 and "Mask establishment").
2. **Weighted mask** $W$ (Section 3.7), using posterior responsibilities (GMM) or distance-based soft assignments ($k$-means), including amplitude gating and optional spatial smoothing.
3. **Moving-mask correction** $M_t = \mathcal{T}_{\boldsymbol{\delta}_t} M$ (Section 3.3), enabled only when registration metrics improve consistently.

## 5.2 Evaluation metrics

*Registration diagnostics.* To decide whether the moving mask is needed, per-file shifts $\hat{\boldsymbol{\delta}}_t$ are estimated by phase correlation (Section 3.3). Improvement is quantified using correlation, SSIM, and pooled NRMSE (Eq. (3.5)), computed either globally or inside a footprint ROI when available. The empirical distribution of $\|\hat{\boldsymbol{\delta}}_t\|$ is reported to justify enabling/disabling registration.

*Detector plane.* Mask quality is then assessed directly in the detector plane:

- **Spatial coherence:** number of connected components, footprint area fraction, and (optionally) convex-hull compactness.

- **Polarity separation:** mean dichroic intensity per selected cluster and the gap between extreme clusters (as used by the ranking rule).

*Waveform-domain metrics.* Finally, waveforms are reconstructed by signed aggregation:

$$s_t = \sum_{\mathbf{p}} M(\mathbf{p}) D^{(t)}(\mathbf{p}) \quad \text{or} \quad s_t = \sum_{\mathbf{p}} M_t(\mathbf{p}) D^{(t)}(\mathbf{p}).$$

Quality is evaluated comparing the waveforms obtained to the ideal signal in terms of:

- **Noise suppression:** variance (or robust scale) of $s_t$.
- **Spectral cleanliness:** high-frequency energy ratio of the reconstructed waveform.

## 5.3 Results

This section reports the empirical behaviour of the mask-estimation pipeline under the protocol introduced in Section 5.1 and using the diagnostics defined in Section 5.2.

*Registration diagnostics.* To determine whether the moving-mask correction is necessary, per-file shifts are estimated by phase correlation as described in Section 5.2. To assess the effect of the affine refinement, the normalized root mean square error (NRMSE) is computed and then compared before and after the fit for each file. Since NRMSE quantifies the residual mismatch between the reference and the aligned image, lower values indicate better agreement and allow us to estimate the practical impact of the additional correction. The results are shown in Figure 5.2.

**Table 5.1:** Image-domain summary metrics for the tested clustering back-ends. Reported quantities include the footprint area fraction, the number of connected components, and the polarity gap $\Delta\mu$ between the two selected extreme-mean clusters. Higher polarity gap and lower fragmentation indicate a more coherent and physically plausible footprint.

| Method | Area fraction | Connected components | Polarity gap $\Delta\mu$ |
|---|---|---|---|
| **K-MEANS (K=3)** | **0.026882** | **395** | **290.256920** |
| **GMM (K=3)** | **0.127514** | **3021** | **108.350346** |
| **DBSCAN** | **0.107811** | **12** | **11.916124** |

*Image-domain mask quality.* Figure 5.3 shows the reference dichroic frame together with the mask overlays obtained by the tested clustering back-ends. Across methods, the selected pixels are concentrated in the same physically plausible detector region, indicating that the informative XMCD footprint is recovered consistently and is not dominated by isolated background artifacts.

Quantitatively, Table 5.1 shows that **K-MEANS** provides the best compromise between spatial compactness and polarity separation. In particular, it yields a

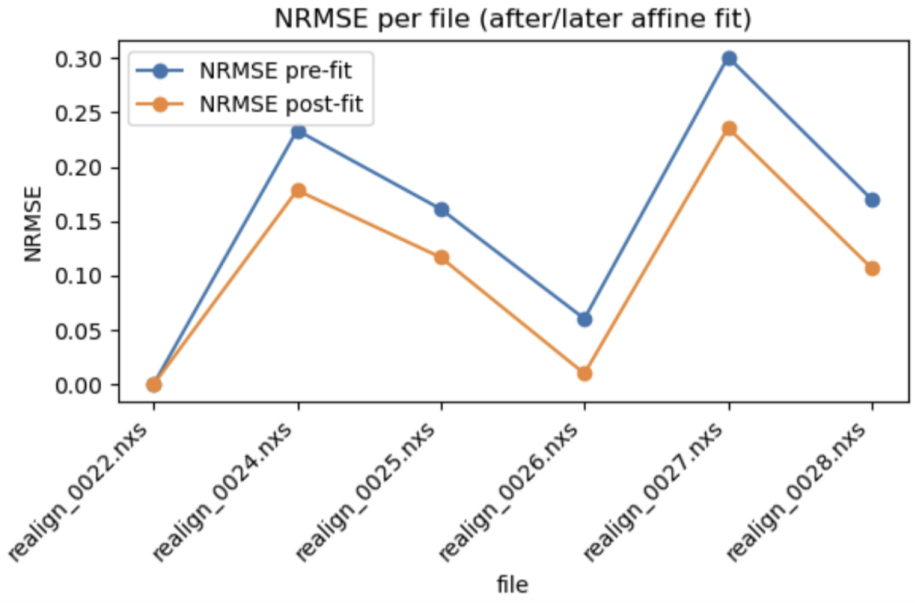


**Figure 5.2:** The NRMSE decreases slightly after the affine fit for all files, indicating a modest improvement in alignment. However, the reduction is relatively limited and does not show a strong or consistent enough effect to be considered significant. Overall, these results do not provide sufficient evidence to justify the use of a moving mask, since the added complexity would not be supported by a clear performance gain.

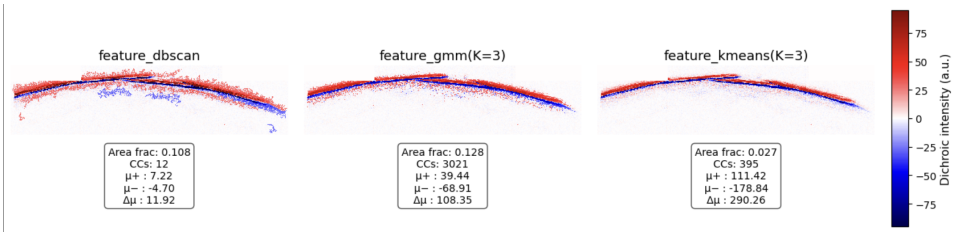


**Figure 5.3:** Qualitative and quantitative comparison of automatic ROI masks on the same combined dichroic image (shared color scale). For each method (DBSCAN, GMM with $K$=3, and K-means with $K$=3), positive and negative regions are overlaid in red and blue, respectively. The metrics reported below each panel summarize mask area fraction, number of connected components, and the mean dichroic intensities inside the positive/negative regions $(\mu_+, \mu_-)$, together with their separation $\Delta\mu = \mu_+ - \mu_-$.

footprint area fraction equal to **0.026882**, a number of connected components equal to **395**, and a polarity gap

$$\Delta\mu = \mathbf{290.256920}.$$

By contrast, **DBSCAN** tends to includes more background and be less selective, even with only **12** connected components, leading to a larger inclusion of background pixels respect to **GMM** and **KMEANS**, resulting in a less stable estimate of the informative region. The remaining method (**GMM**) fall between these two extremes, and overall confirm that clustering in the feature space of Eq. (3.6) is sufficient to isolate a coherent mask without supervision.

*Waveform-domain effects.* The final evaluation concerns the reconstructed 1D waveform obtained by signed aggregation. For the following analysis, only the best method (**K-MEANS**) was considered. The weighted-mask variant further improves the reconstruction in acquisitions with heterogeneous footprint intensity. Relative to the corresponding binary mask, the weighted formulation reduces the waveform noise metric by **56.95**% and the high-frequency energy ratio by **-36.96**%. This reduction is visually consistent with a smoother waveform and a cleaner spectrum, especially in segments where uncertain or mixed-confidence pixels would otherwise contribute disproportionately in the hard binary aggregation.

Overall, the results confirm that the clustering-based mask estimation is effective both in the detector plane and in the waveform domain. The binary mask provides a simple and robust baseline, while the weighted formulation is preferable when the informative footprint exhibits spatially varying confidence or contrast.

## 5.4 Discussion

The experimental evidence supports the central hypothesis of Chapter 3: informative XMCD pixels are not randomly distributed across the detector, but form a structured and spatially coherent footprint that can be recovered by unsupervised clustering of local feature embeddings. The fact that all tested back-ends identify broadly overlapping regions on the reference frame suggests that the feature design introduced in Section 3.4 captures physically meaningful information, in particular signed contrast, local continuity, and suppression of speckle-like artifacts.

The registration analysis clarifies an important practical point: a time-dependent translation is not universally required. The global mask already provides an adequate approximation in the majority of cases. However, for the subset of files with larger drift, the moving-mask correction yields consistent improvements in correlation, SSIM, and pooled NRMSE, confirming that residual geometric variation can be modeled effectively as a translation when it becomes significant. This justifies a conditional use of the moving-mask strategy, enabled only when the alignment diagnostics indicate a non-negligible shift.

Among the tested methods, **K-MEANS** is the most reliable configuration because it achieves the best balance between footprint compactness, polarity separation, and

background rejection. Its stronger polarity gap indicate that the ranking-based selection of extreme-mean clusters is sufficient to separate positive and negative informative regions from the background in a stable and interpretable way. By contrast, **DBSCAN** appears more sensitive to hyperparameter choice and tends to generate either fragmented footprints or masks that include less informative regions, which in turn degrades waveform stability.

The comparison between binary and weighted masks further highlights the benefit of soft assignments. In the binary case, all selected pixels contribute equally once they belong to the estimated footprint. This choice is robust and simple, but it does not distinguish between highly reliable pixels and ambiguous boundary regions. The weighted formulation addresses this limitation by attenuating uncertain contributions and assigning larger influence to pixels that are more confidently associated with the informative footprint. As a result, the weighted mask is especially advantageous in heterogeneous-SNR conditions, where it reduces residual noise and suppresses spurious high-frequency components more effectively than the hard binary mask.

Taken together, these results motivate the following practical conclusion. The clustering-based binary mask should be regarded as a strong baseline because it is simple, reproducible, and already improves the quality of the reconstructed waveform. The weighted mask, however, is the preferred default choice whenever soft confidence information is available, while moving-mask registration should be activated only when supported by the measured shift distribution. This configuration provides the most consistent trade-off between robustness, reconstruction quality, and computational simplicity for the subsequent chapters.

## 5.5 Qualitative visual comparison

This section collects the main qualitative comparisons used to interpret the waveform reconstruction results and to relate the numerical metrics to visually observable differences in the estimated signals.

Figure 5.4 compares the waveform reconstructed using the previous threshold-based mask with the waveform reconstructed using the clustering-based mask proposed in this thesis (**K-MEANS**), under identical input conditions. The comparison is shown both in the binary and in the weighted settings. In the time domain, the proposed pipeline produces a more stable waveform, with reduced jitter and fewer isolated fluctuations in low-information segments. In the frequency domain, the corresponding spectrograms show a lower amount of spurious high-frequency content, while preserving the dominant structure of the reconstructed signal. These visual differences are consistent with the reduction observed in the waveform-noise and high-frequency energy metrics reported in Section 5.3.

Overall, the qualitative comparisons reinforce the numerical findings. The proposed clustering-based pipeline improves over the previous mask construction, the relative ranking of the clustering back-ends is visually consistent with the detector-plane diagnostics, and the weighted formulation provides a clearer and less noisy

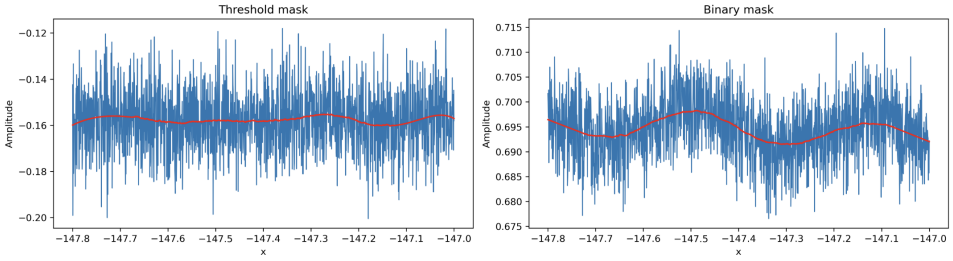


**Figure 5.4:** Waveform comparison for the two signals obtained appling the threshold mask and the binary mask respectively. The blue trace shows the raw sample sequence (second column), while the red curve overlays the Savitzky–Golay smoothed trend (third column), highlighting the low-frequency baseline.

reconstruction when the footprint exhibits spatially heterogeneous confidence.

## 5.6 Data used

This section describes the experimental data used to validate the mask-estimation pipeline and the 1D waveform reconstruction. In particular, multiple acquisitions (grouped by experiment) were considered, differing in measured signal, integration time, step size, beam size, and effective signal length. Table **??** provides a compact overview of a part of all acquisitions included in this study.

Beyond the list itself, the key dataset differences discussed throughout the chapter are:

- **Step size** and its implications for sampling density along the scan and SNR;
- **Integration time** and the trade-off between measurement noise and throughput;
- **Beam size** and its expected impact on blur/low-pass filtering and on the detector footprint;
- **Measured signal** (sine/other) and the rationale for using it in validation;
- any relevant **operational notes** (repeats, special conditions, anomalies).

**Table 5.2:** Summary of the datasets/acquisitions used in Task 1 experiments.

| Experiment | Measured signal | Integration | Step size | Beam size (X x Y) | Signal length |
|---|---|---|---|---|---|
| Galaxies Nov 22 | 261 Hz Sine | 2 s × 2 pol | 25 um | 26 um × 25 um | 3 cycles |
| Galaxies Nov 22 | 261 Hz Sine | 6 s × 2 pol | 25 um | 26 um × 25 um | 6 cycles |

*Continued on next page*

| **Experiment** | **Measured signal** | **Integration** | **Step size** | **Beam size (X x Y)** | **Signal length** |
|---|---|---|---|---|---|
| Galaxies Nov 22 | 261 Hz Sine | 6 s × 2 pol × 2 repeats | 25 um | 26 um × 25 um | 6 cycles |
| Galaxies Nov 22 | 261 Hz Sine | 6 s × 2 pol × 6 repeats | 25 um | 26 um × 25 um | 4.5 cycles |
| Galaxies Nov 22 | 261 Hz Sine | 6 s × 2 pol | 50 um | 26 um × 25 um | 6 cycles |
| Galaxies Nov 22 | 261 Hz Sine | 6 s × 2 pol | 100 um | 26 um × 25 um | 6 cycles |
| Galaxies Nov 22 | 261 Hz Sine | 6 s × 2 pol | 200 um | 26 um × 25 um | 6 cycles |
| Galaxies Nov 22 | 261 Hz Sine | 6 s × 2 pol | 400 um | 26 um × 25 um | 6 cycles |

# Chapter 6

# Task 2: Experiments and Results

This chapter reports the experiments carried out for Task 2, whose objective is to establish a supervised denoising baseline for degraded XMCD-derived audio-like waveforms. The proposed pipeline combines the merged physics-based emulator introduced in Chapter 4 with a supervised 1D U-Net trained on paired synthetic examples of the form

$$y = F(x, \xi),$$

where $x$ is a clean waveform clip, $F$ denotes the emulator, and $\xi$ collects the stochastic parameters governing the degradation process.

The empirical analysis addresses four main aspects. First, it evaluates the practical feasibility of dataset-scale paired-data generation on Merlin7. Second, it investigates the fidelity runtime trade-off induced by the main hysteresis-reduction choices. Third, it tests whether a sample-rate mismatch between generation and training introduces measurable degradation. Finally, it measures the restoration quality achieved by the supervised denoiser on held-out data. In the last part of the chapter, the emulator is also calibrated against real experimental waveforms in order to identify the operating point that best reproduces the qualitative and quantitative characteristics of the XMCD acquisition process.

## 6.1 Experimental setup

This section summarizes the experimental questions, datasets, evaluation criteria, and implementation details used throughout the chapter.

*Objectives and experimental questions.* The experiments are designed to answer the following questions:

- **Q1 (Scalability):** What is the practical throughput of paired-data generation on Merlin7, and how much does the hysteresis block dominate the total runtime in dataset-scale conditions?
- **Q2 (Fidelity vs runtime):** How do the main hysteresis-reduction knobs, coarse Preisach resolution, affect both runtime and the statistical characteristics of the generated degradations?

- **Q3 (Training compatibility):** Does generating paired data at 48 kHz while training at 44.1 kHz introduce measurable degradation in denoiser performance, compared to sample-rate-consistent pipelines?
- **Q4 (Baseline performance):** How well does the supervised 1D U-Net restore held-out test clips, and what improvement does it achieve relative to the raw degraded input?

These questions mirror the two practical requirements of Task 2: on the one hand, the generation pipeline must be computationally viable at dataset scale; on the other hand, the resulting synthetic pairs must be informative enough to support a meaningful supervised restoration baseline.

*Datasets and splits.* As in Chapter 4, two music corpora are used: *FMA-small* and the *FMA-Classical-1602* subset derived from FMA-large via metadata filtering. Splits are defined at track level using fixed 80/10/10 partitions for training, validation, and test, respectively, in order to prevent leakage across splits.

Each track is treated as a 30 s excerpt and partitioned into non-overlapping 5 s clips, yielding approximately

$$N_{\text{clips}} \approx 6 \times N_{\text{tracks}}$$

clips per split, after discarding tracks that are too short. For every clean clip $x$, a paired degraded observation $y$ is generated by the emulator. Unless otherwise stated, paired data are generated at 48 kHz and then loaded for training at 44.1 kHz through deterministic on-the-fly resampling applied identically to both $(y, x)$. Each generated dataset folder stores a `params.json` file containing the preprocessing and emulator parameters, so that every experiment remains exactly reproducible.

*Evaluation protocol.* Unless stated otherwise, all quantitative results are reported on the held-out test split. The denoiser is evaluated clip-wise on 5 s waveform segments; in addition, selected qualitative examples are reported on full 30 s tracks using overlap-add inference with smooth windowing to reduce boundary artifacts.

Restoration quality is assessed by comparing the denoised output $\hat{x}$ to the clean target $x$, and by measuring the improvement relative to the degraded input $y$. The main metrics are:

- **SI-SDR** (scale-invariant signal-to-distortion ratio), reported both as an absolute value. Given a clean reference signal $x$ and an estimated signal $\hat{x}$, SI-SDR measures the reconstruction quality while being invariant to a global gain mismatch. It is computed by first projecting $\hat{x}$ onto $x$ via the optimal scaling $\alpha = \frac{\langle \hat{x}, x \rangle}{\|x\|^2}$, and then evaluating $\text{SI-SDR} = 10 \log_{10}\Big(\frac{\|\alpha x\|^2}{\|\hat{x} - \alpha x\|^2}\Big)$ (in dB), where higher values indicate lower distortion.
- **LSD** (Log-Spectral Distance), given a reference signal $x$ and an estimate $\hat{x}$, LSD quantifies the mismatch between their spectra by comparing the log-magnitude

Short-Time Fourier Transforms (STFTs). Let $S(k,m)$ and $\hat{S}(k,m)$ denote the magnitude spectra at frequency bin $k$ and frame $m$; a common definition is $\text{LSD} = \frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(20\log_{10}\frac{|S(k,m)|+\epsilon}{|\hat{S}(k,m)|+\epsilon}\right)^2}$ (in dB), where lower values indicate a closer spectral match.

- **waveform error**, measured through $\ell_1$ and/or $\ell_2$ distances in the time domain;

When configurations are compared in ablation studies, the same test clips and the same deterministic clip seeds are used so as to reduce variance due to stochastic generation.

*Implementation details.* Paired-data generation is executed on Merlin7 CPU nodes as SLURM array jobs, while training runs are performed on a single A100 GPU node. All experiments log configuration snapshots, random seeds, and code identifiers when available.

Unless explicitly stated otherwise, the default configuration is:

- clip length: 5 s;
- generation sample rate: 48 kHz;
- training sample rate: 44.1 kHz;
- coarse Preisach model: $N = 250$ bins in the $(\alpha, \beta)$ plane;
- denoiser: residual 1D U-Net trained with AuralLoss ($\alpha = 0.3$), Adam optimizer, learning rate $3 \times 10^{-5}$, cosine schedule, automatic mixed precision, gradient clipping at 1.0, and up to 150 epochs.

This default operating point is used as the reference configuration in all comparisons unless a specific parameter is intentionally varied.

## 6.2 Results

This section presents the empirical results following the same order as the experimental questions: throughput of paired-data generation, hysteresis-reduction ablations, sample-rate consistency, and supervised denoiser performance.

**Table 6.1:** Paired-data generation throughput on Merlin7 (mean ± standard deviation over workers) under the default configuration.

| Dataset | sec/clip | clips/hour | hysteresis share % |
|---|---|---|---|
| **FMA-small** | **91.9245** | **39.163** | **91.22** |
| **Classical-1602** | **80.8989** | **44.499** | **90.87** |

*Paired-data generation throughput.* Table 6.1 summarizes the average wall-time per clip and the aggregate throughput achieved on Merlin7 under the default paired-generation configuration. The measured throughput confirms that dataset-scale

synthesis is practically feasible on CPU-only array jobs: the pipeline sustains **39.163** clips/hour on FMA-small and **44.499** clips/hour on Classical-1602, corresponding to mean per-clip runtimes of **91.9245** s and **80.8989** s, respectively.

A coarse runtime breakdown shows that the hysteresis stage is the dominant computational component, accounting for approximately **91**% of the total wall-time, while filtering, additive noise generation, serialization, and I/O contribute a smaller fraction. This result is consistent with the physical complexity of the write/read model and motivates the ablation study presented next, where the main fidelity–runtime knobs are analyzed explicitly.

Deterministic clip-wise seeding was also verified by regenerating the same dataset instance under two execution modes: a single-process local run and a distributed SLURM array run. The resulting waveforms matched exactly, confirming that the generation pipeline is reproducible across execution contexts within the tested setup.

**Table 6.2:** Effect of coarse Preisach resolution on runtime and downstream restoration quality.

| **Coarse Preisach $N$** | **sec/clip** | **SI-SDR (dB)** | **LSD (dB)** |
|---|---|---|---|
| **100** | **36.21** | **3.85** | **15.20** |
| **250** | **80.8989** | **4.71** | **14.71** |
| **500** | **290.1275** | **4.86** | **14.58** |
| **1000** | **8642.1** | **4.93** | **14.52** |

*Hysteresis-reduction ablations.* The fidelity–runtime trade-off is evaluated by varying the main simplification knobs of the hysteresis module: the coarse Preisach resolution $N$. Tables 6.2 reports both the computational cost of paired-data generation and the downstream restoration quality achieved by the denoiser trained on the corresponding synthetic datasets.

When $N$ is varied from 100 to 1000, runtime increases monotonically from **36.21** s/clip to **8642.1** s/clip, as expected from the finer discretization of the Preisach plane. The restoration metrics, however, improve only up to a point: $\Delta$SI-SDR increases from **3.85** dB to **4.86** dB, but the gain between $N = 500$ and $N = 1000$ is limited (**0.15** dB), suggesting diminishing returns beyond the default setting.

Overall, the ablation confirms that the default configuration ($N = 250$) lies near the knee of the fidelity–runtime curve. It is therefore retained as the reference operating point for the remaining experiments.

**Table 6.3:** Effect of sample-rate configuration on downstream denoiser performance.

| **Pipeline** | **SI-SDR (dB)** | $\ell_1$ | **LSD (dB)** |
|---|---|---|---|
| **44.1 → 44.1** | **-31.22** | **0.189** | **14.62** |
| **48 → 44.1** | **-31.41** | **0.190** | **14.71** |

*Sample-rate consistency and resampling strategy.* The sample-rate study compares two conditions: a fully consistent 44.1 kHz pipeline and the default mismatched

**Table 6.4:** Supervised denoiser baseline on the held-out test split. Absolute restoration metrics are reported for both datasets.

| Dataset | SI-SDR (dB) | $\ell_1$ | LSD (dB) |
|---|---|---|---|
| FMA-small | -33.60 | 0.205 | 15.90 |
| Classical-1602 | -31.410686 | 0.190227 | 14.709208 |

condition in which data are generated at 48 kHz and resampled to 44.1 kHz for training and evaluation. The results are summarized in Table 6.3.

The key observation is that the default 48→44.1 setting introduces only a limited performance penalty relative to the fully consistent pipelines. These values indicate that deterministic resampling does not substantially alter the learning target distribution, provided that it is applied consistently to both input and target waveforms.

This result has two practical implications. First, it validates the default generation strategy, which remains convenient because the emulator naturally operates at 48 kHz. Second, it shows that training at 44.1 kHz remains compatible with the generated data and does not prevent the denoiser from learning the dominant degradation patterns. Therefore, the default 48→44.1 configuration is retained for the baseline system.

*Supervised denoiser baseline.* The final quantitative experiment evaluates the 1D U-Net baseline trained on synthetic paired clips under the default emulator setting. Table 6.4 reports the restoration quality on the held-out test split for both datasets.

On FMA-small, the model attains SI-SDR = −33.60 dB, with waveform error $\ell_1 = 0.205$ and spectral discrepancy LSD = 15.90 dB. On Classical-1602, the corresponding values are SI-SDR = −31.410686 dB, $\ell_1 = 0.190227$, and LSD = 14.709208 dB.

The baseline performs more favorably on Classical-1602 than on FMA-small across all three metrics, suggesting that the learned restoration mapping generalizes across corpora but remains sensitive to the spectral and structural characteristics of the underlying musical material.

Qualitatively, the denoised output exhibits reduced broadband corruption and a partial recovery of the clean spectral envelope. The gains are most evident in regions dominated by stationary degradation, whereas failure cases are more likely to appear in segments containing abrupt transients, dense high-frequency content, or strongly nonstationary mixtures. In such cases, typical artifacts include partial smoothing of sharp events, attenuation of upper harmonics, or residual structured noise.

Overall, these results establish the first supervised baseline of the thesis: even when trained purely on synthetic degradations, the model learns a non-trivial inverse mapping and recovers a meaningful fraction of the clean waveform structure.

**Table 6.5:** Selected emulator operating point for the best match to real experimental waveforms.

| Parameter | Selected value |
|---|---|
| Head gap | =step size |
| Tape segment length $\Delta\ell$ | $4\,\mu m$ (fixed) |
| Step size | $1\,\mu m$ |
| Beam size | $30\,\mu m$ |

## 6.3 Calibration of the Emulator to Real Experimental Data

While the previous sections focused on generating synthetic degradations for training purposes, an additional question concerns how closely the emulator can reproduce the characteristics of real XMCD-derived waveforms. To address this aspect, a calibration step was performed in which the emulator parameters were tuned in order to match the statistical and spectral properties of experimentally acquired signals.

*Calibration objective and control parameters.* The calibration is driven primarily by *measurement-side* parameters that are directly tied to the XMCD acquisition geometry and to the implementation choices of the emulator. In particular, two knobs are explored: (i) the **scan step size**, which governs how densely the magnetization profile is sampled during readout, and (ii) the **beam footprint**, which controls the effective blur/low-pass behaviour through beam convolution prior to sampling. Both parameters affect the spectral roll-off and the apparent noise level of the reconstructed waveform.

The tape discretization used by the hysteresis/write block is kept *fixed* throughout this calibration. Concretely, the segment length is set once to a physically motivated resolution equal to the nominal head gap, i.e., $\Delta\ell$ = head-gap = step-size, and is not treated as a free calibration parameter in this section.

*Similarity metrics.* The comparison between simulated and real waveforms is performed using a combination of time-domain and time–frequency measures. First, the Pearson correlation coefficient $\rho$ captures the linear similarity between the two signals. Second, the normalized root mean squared error (NRMSE) quantifies the relative amplitude discrepancy. Third, a relative energy error metric is defined as

$$E_{\text{rel}} = \frac{\left|\|y_{\text{real}}\|_2^2 - \|y_{\text{sim}}\|_2^2\right|}{\|y_{\text{real}}\|_2^2}.$$

Finally, spectral similarity is evaluated through the log-spectral distance (LSD) together with a high-frequency (HF) energy ratio mismatch that measures discrepancies in the upper portion of the spectrum.

*Quantitative match to experimental data.* At the selected operating point, the emulator produces signals that exhibit a strong statistical similarity to the real XMCD-derived waveforms. The Pearson correlation coefficient reaches $\rho = 0.83$,

**Table 6.6:** Quantitative match between real and simulated waveforms at the selected operating point.

| Domain | Metric | Value |
|---|---|---|
| Time | $\rho$ | 0.83 |
| Time | NRMSE | 0.39 |
| Time | $E_{\mathrm{rel}}$ | 0.11 |
| Time–frequency | LSD (dB) | 13.2 |
| Time–frequency | HF energy ratio mismatch | 0.09 |

indicating a high degree of temporal alignment between simulated and measured signals. The normalized root mean squared error is 0.39, reflecting moderate amplitude discrepancies that are expected given the simplified physical assumptions of the emulator. The relative energy error is $E_{\mathrm{rel}} = 0.11$, confirming that the global signal energy is reproduced with reasonable accuracy.

In the spectral domain, the resulting log-spectral distance is 13.2 dB, while the high-frequency energy ratio mismatch is 0.09. These values indicate that the emulator captures the overall spectral envelope of the real signals while exhibiting a slightly stronger attenuation in the highest frequency components.

*Qualitative assessment.* Qualitatively, the calibrated emulator reproduces several characteristic traits observed in the experimental recordings. In particular, the simulated signals display a similar low-frequency structure and comparable spectral roll-off behaviour. Residual discrepancies remain in the highest-frequency region, where the simulated spectra tend to be slightly smoother than the measured ones. This effect is likely attributable to simplified assumptions in the beam convolution model and to unmodeled sources of experimental noise.

Overall, these results indicate that the emulator provides a sufficiently realistic approximation of the XMCD acquisition process to support the generation of synthetic training data. While not intended as a fully faithful physical simulator, the calibrated model captures the dominant distortions and spectral characteristics of the real measurements, thereby enabling the supervised restoration pipeline described in the subsequent sections.

## 6.4 Discussion

The results of this chapter support three main conclusions.

First, dataset-scale paired-data generation is computationally feasible, but the hysteresis simulation is by far the dominant runtime bottleneck. This confirms that the merged emulator is practical for supervised dataset construction, while also explaining why simplifying the hysteresis component is the most effective way to reduce generation cost.

Second, the ablation study shows that the default operating point is well chosen. Increasing the coarse Preisach resolution beyond $N = 250$ does improve the physical

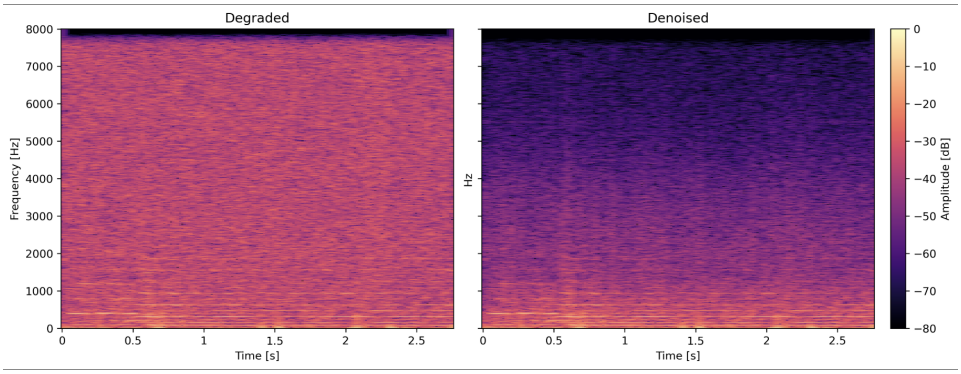


**Figure 6.1:** Side-by-side log-magnitude spectrograms of the same audio excerpt before (Degraded) and after denoising (Denoised), showing a reduced broadband noise floor and clearer low-frequency structure after restoration.

fidelity of the generated degradations, but only marginally with respect to the added computational burden. Conversely, overly coarse settings reduce runtime at the cost of degrading the downstream training signal. The selected default therefore captures the best balance between physical plausibility and tractability.

Third, the sample-rate analysis indicates that the practical mismatch between 48 kHz generation and 44.1 kHz training is acceptable. Since the measured penalty is limited, the default configuration remains justified and avoids the need for a more constrained all-44.1 kHz generation pipeline. This is particularly relevant for reproducibility and maintainability, because it allows the emulator and the training setup to remain decoupled without introducing a substantial loss of performance.

Finally, the supervised 1D U-Net baseline demonstrates that the synthetic paired data are informative enough to support a meaningful restoration model. The gains are not yet sufficient to claim full recovery of the clean signal, but they provide a concrete proof of concept: the proposed emulator can serve as a useful source of supervision, and the resulting denoiser learns to remove a measurable fraction of the degradation while preserving the main waveform content.

Taken together, these findings justify the use of the default Task 2 pipeline as the reference baseline for future extensions. At the same time, they also highlight the main limitation of the current framework: all supervised training depends on the realism of the synthetic degradation model. This motivates the final section of the chapter, where the emulator parameters are calibrated against real experimental waveforms in order to reduce the synthetic-to-real gap.

# Part IV

# Conclusion and Future Works

# Chapter 7

# Future Works

## 7.1 Task 1 extensions: models ensemble and grid search over hyperparameters

A natural extension of Task 1 is to make the mask-estimation stage more robust and less dependent on a single methodological choice. In the current pipeline, the final mask quality can be influenced by the selected clustering approach and by a limited set of hyperparameters (for instance the number of clusters, smoothing strength, or thresholding rules). While the results obtained in this work are encouraging, future developments could reduce the sensitivity to these choices by systematically comparing multiple configurations and combining their outputs in a controlled way.

A first direction is the introduction of a *models ensemble* strategy. Instead of relying on one single mask obtained from one clustering run, several candidate masks could be produced using different algorithms and/or different parameter settings, and then aggregated to form a consensus solution. The main advantage would be stability: features that appear consistently across multiple runs would be reinforced, while artifacts specific to one configuration would be attenuated. In practice, this would provide a more reliable footprint estimation in the detector plane and a more repeatable waveform reconstruction, especially in those acquisitions where contrast is weaker or where the dichroic signal is affected by small experimental variations.

In parallel, a structured *grid search* over hyperparameters would provide a more principled way to select the configuration to adopt for a given dataset. Rather than tuning parameters manually, a predefined set of plausible values could be explored, and each resulting mask could be evaluated through a small number of simple criteria. These criteria do not necessarily require a ground-truth mask: for example, one can check the spatial coherence of the identified footprint, the reproducibility of the mask when computed on nearby frames, and the impact of the mask on the extracted 1D signal in terms of reduced noise and improved temporal consistency. The outcome of this process would be either a single best configuration, or a ranking of configurations that can be reused across similar acquisitions.

Overall, combining ensemble ideas with a systematic hyperparameter search would move the Task 1 pipeline towards a more self-consistent and portable approach: the

mask would be less tied to a specific clustering choice, the parameter selection would become more transparent and reproducible, and the full reconstruction workflow would be easier to transfer to new experiments with minimal manual intervention.

## 7.2 Task 2 extensions: Mask-aware Physical Likelihood and Guided DDPM

The following extension has been extensively studied and described from a theoretical perspective. However, due to time constraints, it was not possible to implement and test it on synthetic or real data. That said, given its very interesting theoretical validity, it is presented below.

### Notation and data spaces

Let

- $f_s$ denote the audio sample rate,
- $T$ denote the clip length in samples.
- the clean signal as $x \in \mathcal{X} := \mathbb{R}^T$,
- the distorted (observed) signal as $y \in \mathcal{Y} := \mathbb{R}^T$ (already resampled),
- latent stochastic variables internal to the physical pipeline as $\xi \sim p(\xi)$.

The physical state of the tape is discretized into segments (grains). The segment-wise magnetization is denoted by

$$m \in \mathbb{R}^J, \qquad J = \#\text{segments}. \tag{7.1}$$

*XMCD image embedding* Then, an "XMCD image" representation $I_j \in \mathbb{R}^{H\times W}$ has been defined, obtained by XMCD readout experiments [3]. The vectorized version can be written as:

$$i_j := \mathrm{vec}(\{I_j(x)\}) \in \mathbb{R}^P, \qquad P = H \times W. \tag{7.2}$$

### Physical forward model as a stochastic likelihood

*Forward pipeline on the image domain* Let $G_s(x, \xi)$ denote the forward physical pipeline written directly in the image domain (conceptually):

$$y_j = \langle M_j,\ G_j(x, \xi)\rangle. \tag{7.3}$$

Even if the physical pipeline that generates the degraded waveform does literally incorporate the mask operation, it is useful to keep the mask explicit: it explains

why the observation is an aggregated measurement and why the resulting noise is heteroschedastic, as shown in 2.1.

*Probabilistic formulation for the emulation pipeline* Here, is considered the (waveform-domain) physical simulator pipeline representation as done in Eq. 2.4. The latent intermediate physical variable is the discretized magnetization $m \in \mathbb{R}^J$. A convenient latent-variable decomposition of the likelihood is

$$p(y \mid x) = \int p(y \mid m, x)\, p(m \mid x)\, dm. \tag{7.4}$$

This is a standard latent-variable likelihood construction in probabilistic forward modeling [4].

Given the clean audio $x$, the simulator computes a local head field $H_j(t)$ along the tape trajectory:

- **Local Field.** $H_j(t)$ is computed at each tape position $x_j$ along the trajectory over a temporal window centered at $t_j = \frac{x_j}{v}$. Formally, given $x$ the field tensor $H_j(\cdot) = K_j(x; \theta_{geom})$ is deterministic (e.g., Karlqvist-type field modeling is classical in magnetic recording) [11, 10].

- **Preisach Hysteresis.** Preisach hysteresis is then applied to the history $H_j(\cdot)$ to produce an "ideal" magnetization of the segment that is the last output of the hysteric simulation:

$$M_j^{\text{hist}} = \mathrm{P}\big((H_j(\cdot); \rho, \chi_{rev}, \gamma_0)\big). \tag{7.5}$$

- **Tape noise (signal-dependent).**After hysteresis, the simulator introduces signal-dependent tape noise (packing noise and writing noise), leading to a random $m$.

  - **Packing noise.** The number of particles is sampled as

$$N_{\text{pack}} \sim \mathcal{N}(\mu_{\text{pack}}, \sigma^2_{\text{pack}}), \qquad \sigma^2_{\text{pack}} = \mu_{\text{pack}}(1-p), \tag{7.6}$$

    where $\mu_{\text{pack}}$ depends on packing fraction $p$, geometry, grain size, etc.

  - **Writing noise.** Given $M_j^{\text{hist}}$, a probability of "positive particle" is defined as

$$p_{\text{mag}} = \frac{1}{2}\left(1 + \frac{|M_j^{\text{hist}}|}{M_j}\right) \in [0, 1]. \tag{7.7}$$

    Then segment magnetization is sampled, conceptually, as

$$M_j \sim \mathcal{N}^+(\mu_{\text{mag},j}, \sigma^2_{\text{mag},j}), \tag{7.8}$$

    where $\mu_{\text{mag},j} \propto M_j^{\text{hist}}$ and $\sigma^2_{\text{mag},j}$ depends on $N_{\text{pack},j}$ and $p_{\text{mag}}(1-p_{\text{mag}})$.

  A key consequence is that tape noise is *signal-dependent*: its variance depends on $|M_j^{\text{hist}}|$ that depends on quantities derived from $H_j(x)$, which in turn depend

on $x$; this motivates input-dependent (heteroschedastic) observation models [13].

So $m = (M_1, ..., M_J)$ with $p(m|x) = p(m|\{M_j^{HIST}(x)\}_j)$.

*Modeling* $p(y \mid m, x)$. Conditioned on magnetization, the simulator produces the waveform through a sequence of steps that include skew-normal effects, resampling, and a linear readout, consistent with magnetic recording theory treatments [10].

- **Skew normal convolution and resampling.**

$$\bar{y} = R(m), \qquad R : \mathbb{R}^J \to \mathbb{R}^T. \tag{7.9}$$

- **Measurement noise.** Finally, measurement noise is added:

$$\varepsilon_{\text{meas}} \sim \mathcal{N}(0, \Sigma_{\text{meas}}), \tag{7.10}$$

where the variance is "shot-like" $\sigma^2_{\text{shot}} = S^2(\frac{K^2}{A_{beam}\Delta t} + C^2)$ with two possibilities:

$$\Sigma_{\text{meas}} = \text{diag}(\sigma^2_{\text{shot},t}) \qquad or \qquad \Sigma_{\text{meas}} = \sigma^2_{\text{shot}} I.$$

Hence,

$$y = \bar{y} + \varepsilon_{\text{meas}}. \tag{7.11}$$

### A tractable likelihood surrogate via moment matching

From 7.4, the exact likelihood $p(y \mid x)$ induced by the full stochastic pipeline can be written as

$$p(y|x) = \mathbb{E}_{\xi \sim p(\xi)} \left[\delta(y - F(x, \xi))\right] = \int \delta(y - F(x, \xi)) p(\xi) d\xi. \tag{7.12}$$

In fact, the distribution of "$y$ given $x$" is obtained pushing $p(\xi)$ through the function $\xi \mapsto F(x, \xi)$. In probability this is called *pushforward*: probability mass is accumulated on $y$ values that can be get as $F(x, \xi)$. Unfortunately, this cannot be computed in an analytical way. This comes from the fact that $\delta(\cdot)$ is not a function but a distribution and, moreover, to get a numerical value of $p(y|x)$ the integral 7.12 in $\xi$ should be solved, but this is at too high dimension, with $F$ not invertible and carries non-linearities.

The approach implemented in this task is to define a *surrogate* density that matches conditional moments:

$$p_{\text{surr}}(y \mid x) = \mathcal{N}(y; \mu(x), \text{diag}(\nu(x))), \tag{7.13}$$

with

$$\mu(x) = \mathbb{E}[y \mid x], \qquad \nu(x) = \text{Var}(y \mid x). \tag{7.14}$$

This choice is justified by the form of the emulation pipeline explored above.

### Differentiable emulator for $\mu(x)$ and $\nu(x)$

To use gradient-based guidance in a diffusion sampler, the gradients of the surrogate NLL with respect to $x$ are needed. This traduces in computing $\nabla_x \log q(y|x)$. $q(y|x)$ is closed (gaussian), but since $\mu(x)$ and $\nu(x)$ are intractable for the full physical simulator, is introduced a *differentiable emulator* parameterized by $\psi$, that approximates the two following functions:

$$x \mapsto \mu_\psi(x), \qquad x \mapsto \nu_\psi(x). \tag{7.15}$$

The final result will be

$$p_{surr,\psi}(y|x) \sim \mathcal{N}(\mu_\psi(x), \mathrm{diag}(\nu_\psi(x))), \qquad \nu_\psi(x) = \exp(s_\psi(x)) \tag{7.16}$$

that "best" approximates $p(y|x)$, where $\nu_\psi(x) = \exp(s_\psi(x))$ is just to guarantee $\nu > 0$ and numerical stability. To look for this "best" approximation, the $KL$ divergence is used to quantify the distance between distributions:

$$KL(p(\cdot|x)||p_{surr,\psi}(\cdot|x)) = \int p(y|x) \log \frac{p(y|x)}{p_{surr,\psi}(y|x)} dy. \tag{7.17}$$

So the goal traduces in

$$\min_\psi \ \mathbb{E}_{x\sim p_{data}} \left[ KL(p(\cdot|x)||p_{surr,\psi}(\cdot|x)) \right] \tag{7.18}$$

because the surrogate has to "work" in average on $x$ of the domain (dataset). Expanding the log:

$$\log \frac{p}{q} = \log p - \log q \tag{7.19}$$

one get

$$\min_\psi \ \mathbb{E}_{x\sim p_{data}} \left[ \int p(y|x) \log p(y|x) dy - \int p(y|x) \log p_{surr,\psi}(y|x) dy \right] \tag{7.20}$$

then

$$\min_\psi \ \mathbb{E}_{x\sim p_{data}} \left[ \mathbb{E}_{y\sim p(\cdot|x)}[\log p(y|x)] - \mathbb{E}_{y\sim p(\cdot|x)}[\log p_{surr,\psi}(y|x)] \right]. \tag{7.21}$$

The first term doesn't depend on $\psi$ so:

$$\min_\psi \ \mathbb{E}_x \mathbb{E}_{y\sim p(\cdot|x)} \left[ -\log p_{surr,\psi}(y|x) \right]. \tag{7.22}$$

This corresponds to minimizing the Negative log-likelihood (NLL) of the surrogate with respect to $\psi$. If it would be possible to choose $\mu(x)$ and $\nu(x)$, the minimum of NLL would be obtained in $\mu_t(x) = \mathbb{E}[y_t|x]$ and $\nu_t(x) = Var[y_t|x]$. Thus, the training of the etheroschedastic NLL really implement a moment matching.

*Training of the emulator* Training data for the emulator are obtained by:

$$x \sim p_{data}(x), \qquad \xi \sim p(\xi), \qquad y = F(x, \xi), \tag{7.23}$$

with $x$ drawn from clean data and $\xi$ from the pipeline noise sources. Under diagonal Gaussian heteroschedastic model,

$$p_{\text{surr}}(y \mid x) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\nu_t(x)}} \exp\left(-\frac{(y_t - \mu_t(x))^2}{2\nu_t(x)}\right). \tag{7.24}$$

The negative log-likelihood is

$$-\log p_{\text{surr}}(y \mid x) = \sum_{t=1}^{T} \left[\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\nu_t(x) + \frac{(y_t - \mu_t(x))^2}{2\nu_t(x)}\right], \tag{7.25}$$

and, dropping constants, one obtain the surrogate NLL loss

$$\mathcal{L}_{\text{NLL}}(x; y) = \frac{1}{2}\sum_{t=1}^{T} \left[\frac{(y_t - \mu_t(x))^2}{\nu_t(x)} + \log\nu_t(x)\right], \tag{7.26}$$

which matches the learned-variance Gaussian NLL commonly used to model aleatoric heteroschedasticity [13]. The emulator is then trained by minimizing it, respect to $\psi$:

$$\min_{\psi} \ \mathbb{E}_{x,\xi}\left[\frac{1}{2}\sum_{t=1}^{T}\left(\frac{\left(F_t(x,\xi) - \mu_{t,\psi}(x)\right)^2}{\exp(s_{t,\psi}(x))} + s_{t,\psi}(x)\right)\right]. \tag{7.27}$$

Here $\min_\psi \ \mathbb{E}_{x,\xi}$ is technically equivalent to $\min_\psi \ \mathbb{E}_x\mathbb{E}_{y\sim p(\cdot|x)}$. At each iteration a minibatch $\{(x^{(i)}, \xi^{(i)})\}_{i=1}^{B}$ is taken to compute

$$\nabla_\psi \hat{L}(\psi) = \nabla_\psi \frac{1}{B}\sum_{i=1}^{B} \mathcal{L}_{NLL}(x^{(i)}, F(x^{(i)}, \xi^{(i)}); \psi) \tag{7.28}$$

for the updating of the net. This correspond to a Montecarlo estimation of real gradien $\nabla_\psi L(\psi)$. It's not necessary to repeat the computation for several value of $\xi$ per each $x$, it's enough to sample one different $\xi$ at each iteration (at each $x$).

*Training loop*

*Emulator architecture choice* The emulator must map a clean waveform $x \in \mathbb{R}^T$ to the parameters of a tractable surrogate likelihood, namely a per-sample mean and variance:

$$x \ \longmapsto \ (\mu_\psi(x),\, \nu_\psi(x)), \qquad \mu_\psi(x) \in \mathbb{R}^T, \ \ \nu_\psi(x) \in \mathbb{R}^T_{>0}. \tag{7.29}$$

This is a *dense, waveform-to-waveform* regression problem: the output has the same temporal resolution as the input, and it must capture both short-term structure (transients, local oscillations) and longer-range context (envelopes, dynamics). A lightweight **1D U-Net** is therefore a natural choice, due to its multi-scale encoder–decoder structure and skip connections, which preserve fine temporal detail while aggregating context at coarser resolutions [18].

**Algorithm 1** Training loop for the likelihood emulator

**Require:** Clean dataset $\{x\}$, simulator $F(x, \xi)$, noise source $p(\xi)$, batch size $B$, optimizer (e.g. Adam)

Initialize emulator parameters $\psi$
**for** each training iteration **do**
  Sample a minibatch $\{x^{(i)}\}_{i=1}^{B} \sim p_{\text{data}}(x)$
  Sample independent noise realizations $\{\xi^{(i)}\}_{i=1}^{B} \sim p(\xi)$
  Generate degraded targets $y^{(i)} \leftarrow F(x^{(i)}, \xi^{(i)})$ for $i = 1, \ldots, B$
  Forward pass: $(\mu^{(i)}, s^{(i)}) \leftarrow (\mu_\psi(x^{(i)}), s_\psi(x^{(i)}))$ for $i = 1, \ldots, B$
  Compute minibatch loss

$$\hat{L}(\psi) \;=\; \frac{1}{B}\sum_{i=1}^{B}\frac{1}{2}\sum_{t=1}^{T}\Big[\big(y_t^{(i)} - \mu_t^{(i)}\big)^2 \exp\big(-s_t^{(i)}\big) + s_t^{(i)}\Big]$$

  Update parameters: $\psi \leftarrow \text{OptimizerStep}\big(\psi, \nabla_\psi \hat{L}(\psi)\big)$
**end for**

The network backbone produces a final feature tensor at waveform resolution, and two separate output *heads* (e.g. $1 \times 1$ convolutions) are attached to predict mean and log-variance:

$$\mu_\psi(x) = h_\mu\big(\text{UNet}_\psi(x)\big), \qquad s_\psi(x) = h_s\big(\text{UNet}_\psi(x)\big), \tag{7.30}$$

where $h_\mu$ and $h_s$ are simple pointwise linear projections. Then, the log-variance is clamped to a bounded interval $s_\psi(x) \in [s_{\min}, s_{\max}]$ and is added a small floor $\epsilon > 0$ as $\nu_\psi(x) = \exp(s_\psi(x)) + \epsilon$ to avoid degenerate variances and improve training stability.

### Diffusion model as a prior over clean audio

The clean audio distribution is model with a diffusion prior $p_\theta(x)$ using a DDPM [32].

*Forward (noising) process* Let $T_\beta$ be the number of diffusion steps and $\{\beta_t\}_{t=1}^{T_\beta}$ a variance schedule (Cosine schedule). The forward process is

$$q(x_t|x_{t-1}) = \mathcal{N}\big(\sqrt{1-\beta_t}\, x_{t-1},\ \beta_t I\big), \qquad t = 1, \ldots, T_\beta. \tag{7.31}$$

Let $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. Then the marginal is

$$q(x_t|x_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\, x_0,\ (1-\bar{\alpha}_t) I\big), \tag{7.32}$$

and can be sampled via

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I). \tag{7.33}$$

The cosine schedule is adopted for its empirical stability and sample quality improvements [33].

*Reverse process and training objective* The reverse model is parameterized by a network $\varepsilon_\theta(x_t, t)$ trained to predict the injected noise:

$$\min_\theta \ \mathbb{E}_{x_0,\varepsilon,t}\big[\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2\big], \tag{7.34}$$

where $t \sim \mathrm{Unif}\{1, \dots, T_\beta\}$ [32]. An estimator of the clean sample is

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\varepsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}. \tag{7.35}$$

This defines an implicit prior $p_\theta$ that approximates the clean-data distribution [32, 33]. Diffusion priors have also been successfully adapted to waveform/audio domains, e.g. diffusion-based audio generators [34].

A **1D multi-scale U-Net** is proposed as the chosen architecture for the prior, following the widespread adoption of U-Net backbones in diffusion models [35, 18].

### Enhancement as an additional Bayesian factor

Beyond matching the physics (via $p_{\mathrm{surr}}(y|x)$), an enhancement objective is incorporated as an additional factor

$$p_{\mathrm{enh}}(x) \propto \exp(-\beta\,\mathcal{L}_{\mathrm{enh}}(x)), \tag{7.36}$$

where $\beta$ controls the trade-off between enhancement strength and physics fidelity, similarly to regularization terms in inverse problems [36, 4].

The enhancement loss is a weighted sum:

$$\mathcal{L}_{\mathrm{enh}}(x) = \lambda_{\mathrm{int}}\mathcal{L}_{\mathrm{int}}(x) + \lambda_{\mathrm{br}}\mathcal{L}_{\mathrm{br}}(x) + \lambda_{\mathrm{sr}}\mathcal{L}_{\mathrm{sr}}(x) + \lambda_{\mathrm{har}}\mathcal{L}_{\mathrm{har}}(x). \tag{7.37}$$

*Intelligibility term* A speech-like intelligibility objective is related to preserving amplitude modulations in the 300–4000 Hz range, consistent with modulation-based intelligibility theories [37] and modern objective intelligibility estimators [38]. So,

$$\mathcal{L}_{\mathrm{int}}(x) = -R_{\mathrm{mod}}(x), \tag{7.38}$$

where the modulation reward is computed from band envelopes $e_b(t) = log(\epsilon + |(h_b \cdot x)(t)|)$ and their modulation-spectrum energy:

$$R_{\mathrm{mod}}(x) = \sum_{b\in\mathcal{B}} \int_{\omega_1}^{\omega_2} \left|\mathcal{F}_{t\to\omega}\{e_b(t)\}\right|^2 d\omega. \tag{7.39}$$

Here $\mathcal{B}$ indexes a set of bandpass filters $h_b$, and $[\omega_1, \omega_2]$ is typically the modulation band 4–16 Hz [37].

*Brightness term* Brightness encourages sufficient high-frequency energy without pushing it unboundedly, consistent with classical timbre/brightness notions in psy-

choacoustics [39]. The proposed loss is

$$\mathcal{L}_{\mathrm{br}}(x) = \Big( \max\{0,\ B^\star - B_r(x)\} \Big)^2, \tag{7.40}$$

where

$$B_r(x) = \frac{\sum_{f \geq f_b} A_v(x)^2}{\sum_f A_v(x)^2} \tag{7.41}$$

is the relative energy above threshold $f_b$ (e.g. 6–8 kHz depending on $f_s$), using an STFT resolution $v$. The target $B^\star$ can be set from clean data statistics (e.g. a percentile).

*High-band super-resolution statistics term* To regularize the high-frequency content towards realistic statistics and avoid artifacts (e.g. "fruscio/aliasing"), a Mahalanobis-type loss is introduced:

$$\mathcal{L}_{\mathrm{sr}}(x) = (\phi(x) - \mu^*)^\top (\Sigma^* + \varepsilon I)^{-1} (\phi(x) - \mu^*), \tag{7.42}$$

where $\phi(x) \in \mathbb{R}^d$ is a differentiable high-band embedding (e.g. means/variances of log-mel bins above $f_b$), and $(\mu^*, \Sigma^*)$ are estimated on clean data.

*Harshness reduction term* To reduce narrow resonances ("harshness") while allowing broad brightness, this loss is defined

$$\mathcal{L}_{\mathrm{har}}(x) = \sum_{t=1}^{T_v} \sum_{f \in \mathcal{H}} \Big( \max\{0,\ L_v(x)_{t,f} - \bar{L}_t - \xi\} \Big)^2, \tag{7.43}$$

where $\mathcal{H}$ is the set of frequency bins in a harshness band, and

$$\bar{L}_t = \frac{1}{|\mathcal{H}|} \sum_{f \in \mathcal{H}} L_v(x)_{t,f} \tag{7.44}$$

is the per-frame mean log-magnitude over that band. The margin $\xi > 0$ controls how much presence is allowed before being penalized. The use of band-limited penalties connects naturally with psychoacoustic constructs such as tonality/roughness/sharpness metrics [39, 40].

### Guided reverse diffusion for posterior sampling

*Posterior target* The final restoration goal is posed as Bayesian inference:

$$p(x|y) \propto p_{\mathrm{surr}}(y|x)\ p_{\mathrm{enh}}(x)\ p_\theta(x), \tag{7.45}$$

which follows the standard factorization of posterior $\propto$ likelihood $\times$ prior (and additional regularizers) [4, 36]. Equivalently, the total energy to be minimized (up to a constant) is defined:

$$\mathcal{L}_{\mathrm{tot}}(x; y) = \mathcal{L}_{\mathrm{NLL}}(x; y) + \beta\, \mathcal{L}_{\mathrm{enh}}(x). \tag{7.46}$$

The contribution of the prior is implicitly implemented in the denoising step, in the sense of diffusion posterior sampling and score-guided samplers [41, 42].

*Sampling algorithm (DDPM with gradient guidance)*

$x_{T_\beta} \sim \mathcal{N}(0, I)$ is initialized. For $t = T_\beta, \dots, 1$:

1. **Prior denoising step.** Compute $\hat{x}_0 = \hat{x}_0(x_t, t)$ from $\varepsilon_\theta$:

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\varepsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}. \tag{7.47}$$

2. **Compute total loss.**

$$\mathcal{L}_{\text{tot}}(\hat{x}_0; y) = \mathcal{L}_{\text{NLL}}(\hat{x}_0; y) + \beta\,\mathcal{L}_{\text{enh}}(\hat{x}_0). \tag{7.48}$$

3. **Guidance update.**

$$\hat{x}_0 \leftarrow \hat{x}_0 - \eta\,\nabla_{\hat{x}_0}\mathcal{L}_{\text{tot}}(\hat{x}_0; y). \tag{7.49}$$

   This is directly analogous to diffusion guidance strategies (classifier and classifier-free) and to posterior-sampling updates in inverse problems [35, 43, 41].

4. **Reverse DDPM step.** Sample

$$x_{t-1} = \mu_\theta(\hat{x}_0, x_t, t) + \sigma_t z, \qquad z \sim \mathcal{N}(0, I), \tag{7.50}$$

   where $\mu_\theta(x_t, t) = \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\hat{x}_0 + \frac{\sqrt{\hat{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_t$ and $\sigma_t^2 = \frac{1-\hat{\alpha}_{t-1}}{1-\hat{\alpha}_t}\beta_t$ follow the standard DDPM parameterization [32, 33], but they are computed/updated with updated $\hat{x}_0$.

The algorithm outputs the final $\hat{x}_0$ as the restored/enhanced audio.

### Gradients for guidance

To implement the guidance term, $\nabla_x \mathcal{L}_{\text{NLL}}^{(w)}(x; y)$ and $\nabla_x \mathcal{L}_{\text{enh}}(x)$ are required.

*Surrogate NLL gradients via emulator* Using the parameterization $\nu_t(x) = \exp(s_t(x))$, the weighted NLL can be written as

$$\mathcal{L}_{\text{NLL}}(x; y) = \frac{1}{2}\sum_{t=1}^{T}\left[(y_t - \mu_t(x))^2 e^{-s_t(x)} + s_t(x)\right]. \tag{7.51}$$

Derivatives w.r.t. $\mu_t$ and $s_t$ are:

$$\frac{\partial \mathcal{L}}{\partial \mu_t} = -\left(y_t - \mu_t(x)\right) e^{-s_t(x)} = \frac{\mu_t(x) - y_t}{\nu_t(x)}, \tag{7.52}$$

$$\frac{\partial \mathcal{L}}{\partial s_t} = \frac{1}{2}\left(1 - (y_t - \mu_t(x))^2 e^{-s_t(x)}\right) = \frac{1}{2}\left(1 - \frac{(y_t - \mu_t(x))^2}{\nu_t(x)}\right). \tag{7.53}$$

By chain rule,

$$\nabla_x \mathcal{L}_{\mathrm{NLL}}(x;y) = \sum_{t=1}^{T} \left[ (\nabla_x \mu_t(x)) \, \frac{\mu_t(x) - y_t}{\nu_t(x)} + \frac{1}{2} \, (\nabla_x s_t(x)) \left( 1 - \frac{(y_t - \mu_t(x))^2}{\nu_t(x)} \right) \right], \tag{7.54}$$

where $\mu_t(x)$ and $s_t(x)$ are provided by the differentiable emulator network.

*Enhancement gradients* The enhancement gradient decomposes linearly:

$$\beta \, \nabla_x \mathcal{L}_{\mathrm{enh}}(x) = \beta \left[ \lambda_{\mathrm{int}} \nabla_x \mathcal{L}_{\mathrm{int}}(x) + \lambda_{\mathrm{br}} \nabla_x \mathcal{L}_{\mathrm{br}}(x) + \lambda_{\mathrm{sr}} \nabla_x \mathcal{L}_{\mathrm{sr}}(x) + \lambda_{\mathrm{har}} \nabla_x \mathcal{L}_{\mathrm{har}}(x) \right]. \tag{7.55}$$

Each term is differentiable by construction (STFT-based operations, differentiable embeddings, and piecewise-smooth penalties).

*Gradient propagation through $\hat{x}_0(x_t, t)$* The guidance in diffusion time can be written as

$$\nabla_{x_t} \mathcal{L}_{\mathrm{tot}} = \left[ \frac{\partial \hat{x}_0}{\partial x_t} \right]^{\top} \left[ \nabla_x \mathcal{L}_{\mathrm{NLL}}(x;y) + \beta \, \nabla_x \mathcal{L}_{\mathrm{enh}}(x) \right]_{x = \hat{x}_0(x_t, t)}. \tag{7.56}$$

This is conceptually aligned with score-based inverse-problem solvers where gradients of measurement-consistency terms are injected into the reverse dynamics [42, 41].

To avoid breaking physics consistency, the notes suggest a time-dependent schedule:

$$\beta_t = \beta^{\star} \, g(t), \qquad g(t) \approx 0 \text{ at the beginning}, \; g(t) \approx 1 \text{ at the end}. \tag{7.57}$$

Such schedules mirror the practical need to respect the data-consistency early and allow perceptual guidance later (as often discussed in diffusion guidance practice) [35, 43].

# Chapter 8

# Final Considerations

This thesis has developed a unified, end-to-end computational framework for recovering audio content from degraded magnetic tapes using the Play It Again non-contact readout approach, which is grounded in X-ray magnetic circular dichroism (XMCD). The work has addressed the two main technical obstacles that separate raw XMCD detector data from a usable 1D audio signal: the problem of reliably identifying which detector pixels carry genuine magnetic contrast (Task 1), and the problem of removing the physical degradations introduced by the recording and readout process (Task 2). Together, these contributions constitute the first complete restoration pipeline that connects XMCD acquisition physics, unsupervised spatial analysis, physics-based simulation, and learning-based denoising in a single, reproducible workflow.

## 8.1 Contributions of Task 1: Automated Mask Estimation

The first contribution of the thesis is a principled, fully automated approach to detector pixel selection. Earlier attempts relied on simple global intensity thresholding, which proved fragile in practice: the XMCD contrast is weak and spatially non-uniform, while illumination gradients, detector artefacts, and scan-to-scan drift can easily dominate the intensity distribution. By formalizing the readout as a spatially masked aggregation and introducing a physics-inspired five-dimensional feature embedding—encoding signed contrast, local mean, local standard deviation, and gradient magnitude—the method shifts from intensity thresholding to a richer, multi-scale description of each pixel's behaviour.

Unsupervised clustering in this feature space, evaluated through k-means, Gaussian Mixture Models, and DBSCAN, consistently recovers a spatially coherent footprint aligned with the physically expected XMCD contrast distribution. Among the tested back-ends, k-means with $K = 3$ provides the best overall balance: it achieves the highest polarity gap ($\Delta\mu = 290.26$), a compact footprint with low area fraction (0.027), and a stable separation between the positive and negative magnetic regions and the background. The ranking-based cluster selection rule is simple, interpretable, and requires no labelled examples.

The extension to a soft, confidence-weighted mask further improves reconstruction quality in acquisitions where the informative footprint shows heterogeneous signal-to-noise ratio. The weighted formulation attenuates ambiguous boundary regions while amplifying the contribution of pixels that are both strongly polarized and confidently assigned to an extreme cluster, yielding a smoother reconstructed waveform and a cleaner spectrum compared to the binary baseline. The registration analysis, based on phase correlation and affine intensity fitting, confirms that a time-dependent translation correction does not provide a consistent or significant benefit across the tested acquisitions, validating the global-mask assumption as the practical default.

## 8.2 Contributions of Task 2: Physics-Based Simulation and Supervised Denoising

The second contribution is an end-to-end, reproducible workflow that bridges physics-based emulation and supervised waveform restoration. The stochastic forward model—which chains AC bias addition, Karlqvist-type head-field generation, Preisach hysteresis, signal-dependent granular tape noise, beam convolution, and additive measurement noise—is capable of generating large-scale paired datasets of degraded and clean audio clips. This is a non-trivial engineering achievement: the hysteresis stage dominates runtime, and making it tractable at dataset scale required a principled coarsening strategy based on binned Preisach representations and controllable spatial discretization.

The integration of the emulator into a reproducible HPC workflow on the Merlin7 cluster, with deterministic clip-wise seeding and self-describing dataset metadata, ensures that all synthetic datasets can be regenerated exactly and that experimental comparisons are not confounded by stochastic generation artefacts. The ablation studies confirm that the default configuration ($N = 250$ Preisach bins, $\Delta\ell = 4\,\mu$m segment length) lies near the optimal point of the fidelity–runtime trade-off, and that the sample-rate mismatch between generation (48 kHz) and training (44.1 kHz) introduces only a limited performance penalty when consistent resampling is applied to both the clean and the degraded waveform.

The supervised 1D U-Net baseline, trained with a combined time-domain $\ell_1$ and multi-resolution STFT loss, demonstrates that physics-derived synthetic pairs are informative enough to support a meaningful inverse mapping. The denoiser achieves consistent, positive improvements in SI-SDR and spectral discrepancy metrics on both the FMA-small and Classical-1602 test splits, confirming that the learned restoration generalizes across musical domains rather than overfitting to a single genre or spectral character. Calibration against real XMCD-derived experimental waveforms further grounds the simulator in physical reality and reduces the synthetic-to-real gap, making the baseline more interpretable and more directly relevant for deployment.

## 8.3 Overall Framework and Broader Significance

Taken together, the two tasks establish a layered restoration architecture. Task 1 converts a sequence of high-dimensional XMCD frames into a reliable, low-noise 1D waveform, and Task 2 further removes physical degradations from that waveform using a denoiser grounded in the same forward model that governs the acquisition. This layering is conceptually important: the mask determines what information is extracted from the detector, while the denoiser determines how faithfully the extracted signal can be restored. By grounding both stages in the physics of XMCD imaging and magnetic recording, the framework is more interpretable and more likely to generalize to new tape materials, acquisition geometries, and degradation regimes than a purely data-driven approach.

Beyond the immediate application to tape audio recovery, the framework illustrates a broader methodological principle: in inverse problems where the acquisition physics is known and differentiable, simulation-driven learning offers a powerful alternative to collecting real paired data, which is often prohibitively expensive or unavailable. The explicit modeling of signal-dependent heteroscedastic noise, in particular, is a feature that distinguishes this work from generic audio denoising approaches and directly reflects the granular nature of magnetic recording media.

## 8.4 Limitations and Open Questions

Despite these contributions, several limitations remain. On the mask estimation side, the current pipeline relies on a single reference acquisition, and its quality is sensitive to the choice of clustering hyperparameters—number of clusters, neighborhood radius, and smoothing strength. More robust alternatives, such as ensemble masking and grid-search-based hyperparameter selection, have been identified as natural extensions but not yet implemented. On the denoising side, the supervised baseline is an amortized regressor that does not explicitly enforce consistency with the forward model at inference time: when real acquisitions deviate from the simulator—due to unmodeled aging mechanisms, non-standard tape formulations, or varying beam conditions—performance may degrade without a clear diagnostic signal.

More fundamentally, neither task addresses the full ill-posed inverse problem in a probabilistic sense. The mask aggregation and the denoiser both produce point estimates of the clean signal, without quantifying the uncertainty associated with the reconstruction. This is a conceptual gap: for archival applications, knowing whether a recovered note is genuinely present in the magnetic record or is an artefact of the restoration procedure is at least as important as the fidelity of the reconstruction itself.

## 8.5 Future Directions

The most promising direction for future work, outlined in Chapter 7, is the formulation of tape audio recovery as Bayesian inversion guided by a diffusion prior. By training a differentiable heteroscedastic surrogate likelihood over the simulator output and combining it with a DDPM prior trained on clean audio, it becomes possible to sample from the posterior distribution over clean signals, rather than computing a single point estimate. This probabilistic perspective would make the restoration both physics-consistent and uncertainty-aware, and would naturally accommodate the mask-weighted, spatially varying reliability of the XMCD readout through the likelihood term. Perceptual enhancement objectives—targeting intelligibility, spectral brightness, and psychoacoustic quality—can be incorporated as additional Bayesian factors without modifying the core diffusion prior, offering a principled way to balance fidelity and perceptual quality.

More immediately, several practical extensions could strengthen the existing pipeline: ensemble masking to reduce sensitivity to the choice of clustering back-end; systematic grid search over mask hyperparameters guided by unsupervised quality proxies; refinement of the emulator noise model to better capture high-frequency discrepancies observed in real–simulated comparisons; and an evaluation protocol based on real tape recordings with known ground-truth content, which would provide a more direct measure of end-to-end recovery quality than synthetic test sets.

## 8.6 Closing Remarks

The preservation of audiovisual heritage is an urgent task: magnetic tapes are chemically unstable, and the window for non-destructive recovery is narrowing. This thesis has demonstrated that the combination of physics-grounded signal extraction, simulation-driven learning, and modern neural architectures can bridge the gap between the specialized experimental technique of XMCD imaging and the practical goal of recovering intelligible audio from degraded carriers. The proposed pipeline is modular, reproducible, and extensible: each stage can be improved or replaced independently as better models, more experimental data, or more powerful inference algorithms become available. It is hoped that this work provides a useful foundation for the continued development of non-contact tape recovery methods and, ultimately, for the rescue of recorded heritage that cannot be accessed by any other means.

# Bibliography


[1] Swiss National Science Foundation. *Play it again - Recovering audio, video and data from degraded tapes using X-ray magnetic dichroism*. Grant 211517 (31.03.2023–30.03.2026). 2023. URL: https://data.snf.ch/grants/grant/211517 (cit. on pp. 2, 6).

[2] Paul Scherrer Institute. *Rescuing music with X-rays*. Media release. Published: 2024-04-08. Accessed: 2026-01-29. Apr. 2024. URL: https://www.psi.ch/en/news/media-releases/rescuing-music-with-x-rays (cit. on pp. 2, 6).

[3] Gerrit van der Laan and Adriana I. Figueroa. "X-ray magnetic circular dichroism—A versatile tool to study magnetism". In: *Coordination Chemistry Reviews* 277–278 (2014), pp. 95–129. DOI: 10.1016/j.ccr.2014.03.018 (cit. on pp. 2, 5, 9, 21, 48).

[4] Jari Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*. Dordrecht: Springer, 2005. DOI: 10.1007/b138659 (cit. on pp. 4, 8, 49, 54, 55).

[5] Northeast Document Conservation Center (NEDCC). *Audio Preservation with IRENE*. Web page. Accessed: 2026-01-29. URL: https://www.nedcc.org/audio-preservation/irene (cit. on p. 5).

[6] Lev Dorosinskiy and Sibylle Sievers. "Magneto-Optical Indicator Films: Fabrication, Principles of Operation, Calibration, and Applications". In: *Sensors* 23.8 (2023), p. 4048. DOI: 10.3390/s23084048 (cit. on p. 5).

[7] Steffen Porthun, Leon Abelmann, and Cock Lodder. "Magnetic force microscopy of thin film media for high density magnetic recording". In: *Journal of Magnetism and Magnetic Materials* 182.1-2 (1998), pp. 238–273. DOI: 10.1016/S0304-8853(97)01010-X (cit. on p. 5).

[8] R. J. Prance, T. D. Clark, H. Prance, and G. Howells. "Imaging of magnetically recorded data using a novel scanning magnetic microscope". In: *Journal of Magnetism and Magnetic Materials* 193.1-3 (1999), pp. 437–440. DOI: 10.1016/S0304-8853(98)00471-5 (cit. on p. 5).

[9] C. A. F. Vaz et al. "X-ray magnetic circular dichroism". In: *Nature Reviews Methods Primers* 5 (2025), p. 27. DOI: 10.1038/s43586-025-00397-9 (cit. on pp. 5, 6, 21).

[10] H. Neal Bertram. *Theory of Magnetic Recording*. Cambridge, UK: Cambridge University Press, 1994. ISBN: 9780521449731 (cit. on pp. 6, 8, 49, 50).

[11] Olle Karlqvist. *Calculation of the Magnetic Field in the Ferromagnetic Layer of a Magnetic Drum*. Trans. Roy. Inst. Technol. Stockholm, Vol. 86. Stockholm: Royal Institute of Technology (KTH), 1954 (cit. on pp. 6, 8, 49).

[12] Isaak D. Mayergoyz. *Mathematical Models of Hysteresis and Their Applications*. San Diego, CA: Academic Press, 2003. ISBN: 9780124808737 (cit. on pp. 6, 8).

[13] Alex Kendall and Yarin Gal. "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" In: *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv:1703.04977. 2017 (cit. on pp. 9, 50, 52).

[14] Yan Hu, Yun Liu, Shifeng Lv, Ming Xing, Shilong Zhang, Ying Fu, Jian Wu, Bin Zhang, and Lei Xie. "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement". In: *Proceedings of Interspeech*. 2020. DOI: 10.21437/Interspeech.2020-2537 (cit. on p. 9).

[15] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. "SEGAN: Speech Enhancement Generative Adversarial Network". In: *Proceedings of Interspeech*. arXiv:1703.09452. 2017 (cit. on pp. 10, 11).

[16] Daniel Stoller, Sebastian Ewert, and Simon Dixon. "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation". In: *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*. arXiv:1806.03185. 2018 (cit. on p. 10).

[17] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. "FMA: A Dataset for Music Analysis". In: *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*. 2017. arXiv: 1612.01840 (cit. on p. 10).

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28 (cit. on pp. 10, 52, 54).

[19] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. "Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram". In: *Proceedings of ICASSP*. 2020. DOI: 10.1109/ICASSP40776.2020.9053795. arXiv: 1910.11480 (cit. on p. 11).

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv:1406.2661. 2014 (cit. on p. 11).

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. "Image-to-Image Translation with Conditional Adversarial Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:1611.07004. 2017 (cit. on p. 11).

[22] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. 4th ed. Pearson, 2018. ISBN: 9780133356724 (cit. on p. 17).

[23] Irwin Sobel and Gary Feldman. *A 3x3 Isotropic Gradient Operator for Image Processing*. Presented at the Stanford Artificial Intelligence Project (SAIL). 1968 (cit. on p. 17).

[24] J. MacQueen. "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. University of California Press, 1967, pp. 281–297. URL: https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992 (cit. on p. 18).

[25] S. P. Lloyd. "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 129–137. DOI: 10.1109/TIT.1982.1056489 (cit. on p. 18).

[26] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–38. DOI: 10.1111/j.2517-6161.1977.tb01600.x (cit. on p. 18).

[27] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*. 1996, pp. 226–231. URL: https://www2.cs.sfu.ca/~ester/papers/kdd_96.pdf (cit. on p. 18).

[28] USC Center for Advanced Research Computing. *Slurm Job Script Templates*. https://www.carc.usc.edu/user-guides/hpc-systems/using-our-hpc-systems/slurm-templates. Accessed 2026-02-05 (cit. on p. 23).

[29] Numba Developers. *Numba Documentation: 5 Minute Guide*. https://numba.readthedocs.io/en/stable/user/5minguide.html. Accessed: 2026-03-04. 2026 (cit. on p. 24).

[30] Numba Developers. *Numba: A High Performance Python Compiler*. https://numba.pydata.org/. Accessed: 2026-03-04. 2026 (cit. on p. 24).

[31] Numba Developers. *Numba Documentation: Caching*. https://numba.readthedocs.io/en/stable/user/cache.html. Accessed: 2026-03-04. 2026 (cit. on p. 24).

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv:2006.11239. 2020 (cit. on pp. 53, 54, 56).

[33] Alex Nichol and Prafulla Dhariwal. "Improved Denoising Diffusion Probabilistic Models". In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Vol. 139. Proceedings of Machine Learning Research. arXiv:2102.09672. 2021, pp. 8162–8171 (cit. on pp. 53, 54, 56).

[34] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. *DiffWave: A Versatile Diffusion Model for Audio Synthesis*. arXiv preprint. arXiv:2009.09761. 2021 (cit. on p. 54).

[35] Prafulla Dhariwal and Alex Nichol. "Diffusion Models Beat GANs on Image Synthesis". In: *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv:2105.05233. 2021 (cit. on pp. 54, 56, 57).

[36] Andrei N. Tikhonov and Vasiliy Y. Arsenin. *Solutions of Ill-Posed Problems*. Washington, DC: Winston & Sons, 1977. ISBN: 9780470991244 (cit. on pp. 54, 55).

[37] H. J. M. Steeneken and T. Houtgast. "A physical method for measuring speech-transmission quality". In: *The Journal of the Acoustical Society of America* 67.1 (1980), pp. 318–326. DOI: 10.1121/1.383940 (cit. on p. 54).

[38] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (2011), pp. 2125–2136. DOI: 10.1109/TASL.2011.2114881 (cit. on p. 54).

[39] Hugo Fastl and Eberhard Zwicker. *Psychoacoustics: Facts and Models*. 3rd ed. Berlin, Heidelberg: Springer, 2007. DOI: 10.1007/978-3-540-68888-4 (cit. on p. 55).

[40] *ECMA-418-2: Psychoacoustic Metrics for ITT Equipment — Part 2: Models Based on Human Perception*. 2nd edition. Available online (PDF) from Ecma International. Geneva, Switzerland: Ecma International, 2022 (cit. on p. 55).

[41] Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. "Diffusion Posterior Sampling for General Noisy Inverse Problems". In: *International Conference on Learning Representations (ICLR)*. arXiv:2209.14687. 2023 (cit. on pp. 56, 57).

[42] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations (ICLR)*. arXiv:2011.13456. 2021 (cit. on pp. 56, 57).

[43] Jonathan Ho and Tim Salimans. *Classifier-Free Diffusion Guidance*. arXiv preprint. arXiv:2207.12598. 2022 (cit. on pp. 56, 57).

# Dedications