ALF: SPECTRALLY ANCHORED LATENT FLOW MATCHING FOR WIRELESS SIGNAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Wireless time-series generation is useful for emerging applications that will drive the adoption and integration of machine learning tasks within next-generation networks, such as waveform classification in shared spectrum bands and interpretation of the physical world through integrated sensing and communications. Unlike a generic time series, wireless signals have time-frequency duality, where a frequency domain bandwidth constraint implicitly imposes latent structural constraints on the evolution of the time-domain sequence. In this paper, we develop Anchored Latent Flow (ALF), a latent space flow model for label-conditioned signal generation that exploits this duality. Our flow model "reasons" through a spectrogram, which is a 2-D representation of the signal constructed by applying a windowed Fast Fourier Transform to successive, overlapping signal segments and mapping their frequency-domain power components across time. Specifically, we shape the latent space of the flow model through a training-time auxiliary objective that is supervised by the (latent) spectrogram (i.e. anchoring through the spectrogram). By guiding the flow field to reason through a spectrogram, the generated sequence maintains long-range coherence across time and greatly improves the signal quality at inference. Empirically, ALF exhibits state-of-the-art generation quality and reconstruction error across various signal lengths, waveform classes, signal-to-noise-ratio (SNR) and channel conditions, reaching a train on synthetic, test on real (TSTR) accuracy of 84.5%, a $1.9\times$ average memory reduction, and $100\times$ average compute speedup over existing models. Our design further enables downstream tasks including style transfer and data augmentation, and has broad utility for the wireless community.

1 INTRODUCTION

Generating a received radio-frequency (RF) transmission under certain propagation environments and signal-to-noise ratios without conducting an over-the-air experiment is challenging. Such generative capabilities enable applications ranging from targeted synthetic data augmentation to creating digital twins of specific environments. Recent progress in diffusion and flow modeling has helped solve similar problems in the image and video domains (Ho et al., 2020; 2022; Lipman et al., 2023; Liu et al., 2023b). RF signals, however, possess highly engineered structures that demand precise modeling for trustworthy generation. Longer samples, spectral and phase coherence, and correlations across multiple receive antennas all introduce additional challenges.

RF waveforms are naturally constrained in both the time and frequency domain. For example, regulatory and task-specific bandwidth limitations restrict RF signals to a prescribed range of frequencies (Federal Communications Commission, 2026). Due to time-frequency duality, these spectral constraints induce strong structure in the time domain, restricting plausible waveforms to a small subset of the waveform subspace. Additionally, ignoring the spectral structure can produce time series generations that violate spectral constraints or corrupt waveform morphology. Prior work has exploited spectral information through Fourier-based architectures and training objectives (Wu et al., 2023; Yuan & Qiao, 2024) and joint time-frequency diffusion for RF signals (Chi et al., 2024). However, prior work does not leverage the frequency domain as an intermediate reasoning space. We hypothesize that supervising the frequency-domain latent can improve coherence while enabling substantially longer RF generation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

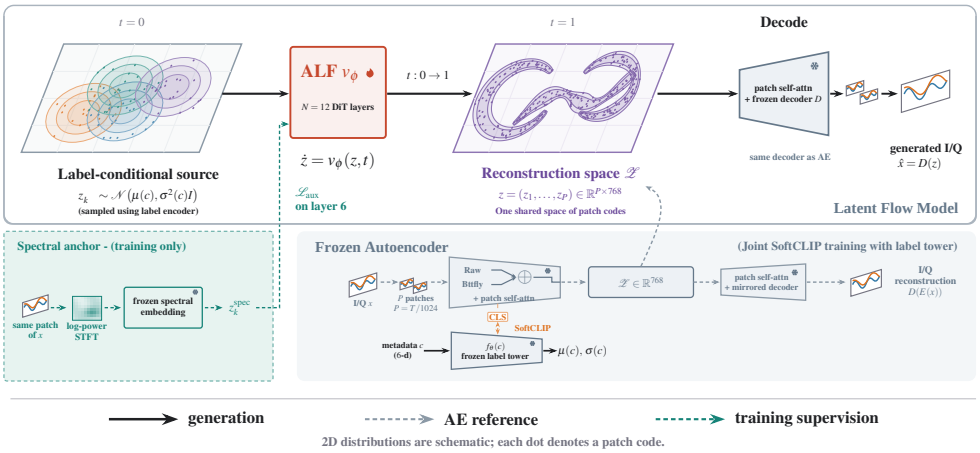


Figure 1: Anchored Latent Flow (ALF) pipeline. A contrastively trained autoencoder learns to reconstruct signals while maintaining alignment with feature labels (§ 5.1). ALF then transports a multimodal feature space to the reconstruction space by reasoning through the spectrogram (§ 5.2).

Resolving how exactly to incorporate spectral information into the generative learning procedure is therefore a critical design choice. We take inspiration from *anchoring* introduced by Rout et al. (2025a) for diffusion language modeling. The central idea is to reveal the most informative tokens (anchors) first, as they guide the remaining generation. Accordingly, the space of anchors resides in the same space as the generated output. In the RF domain, however, individual signal samples carry far less global context to exploit than an important text token would. To this end, our key observation is that "anchors" need not occupy the same subspace as the final generated time-domain output. Instead, we anchor from the frequency-domain representation to shape the latent space of a flow. Specifically, we posit that learning to predict a compact spectrogram can effectively help guide the generation of time-domain data. **Our core contributions are:**

1. We propose **Anchored Latent Flow** (ALF, see Figure 1), a label-conditioned latent flow-matching model for single-input-multiple-output (SIMO) in-phase and quadrature (IQ) generation. ALF transports an aligned, condition-aware signal latent space to a reconstruction-oriented latent space using a learned velocity field. For long-range coherence, we anchor the flow in the frequency domain by supervising the intermediate representation of the flow model to predict spectrogram latents mixed with global information (Appendix A for a formal treatment).
2. We introduce **WaveScene**, a large-scale SIMO dataset containing 7.82 million IQ recording-label pairs across various waveform classes, propagation environments, and sequence lengths. Experiments on unseen channel realizations demonstrate competitive representation learning, long-sequence generation, and controlled signal editing.
3. We show that spectrogram anchoring reshapes the flow's intermediate representations and improves generation over matched unanchored ablations. Overall, ALF achieves a 38.02% relative improvement in Train on Synthetic, Test on Real (TSTR) waveform classification accuracy compared to the closest baseline. We further demonstrate computational speedups by $100\times$, memory reductions by $1.9\times$, native style transfer, and autoregressive capabilities.

2 Related Work

We describe the related work below, deferring a more detailed exposition to Appendix B.

Diffusion Models have emerged as a leading approach to generative modeling for images (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021), audio (Kong et al., 2021), and video (Ho et al., 2022). More recent work has applied diffusion models to time-series tasks, including forecasting (Rasul et al., 2021; Yan et al., 2021; Biloš et al., 2023; Kollovieh et al., 2023; Li et al., 2022) and imputation (Tashiro et al., 2021; Alcaraz & Strodthoff, 2023a; Yuan & Qiao, 2024), using both conditional formulations and guidance-based approaches. Conditional generation of time-series has

been developed in Alcaraz & Strodthoff (2023b), Narasimhan et al. (2024), and Narasimhan et al. (2025). However, our experiments (see Table 2) show that these state-of-the-art (SOTA) approaches struggle to generate long-range coherent IQ waveforms, which are typically required in the RF domain. **Our key hypothesis** is that these approaches *lack supervision of intermediate representations pertaining to the global frequency domain*, and therefore generate sub-optimal results. Alongside conditioning and guidance, recent work in the language domain introduces *anchoring*, in which the model predicts important intermediate information that greatly reduces the entropy of the remaining data to be generated (Rout et al., 2025a). Similarly, representations have been shaped in the image domain using a clean image-target in (Yu et al., 2025). In light of these, we introduce cross-domain spectral anchoring, in which intermediate model representations are supervised using a separately learned frequency-domain representation so that RF-domain-specific characteristics are used as first-class inductive biases.

RF-Specific Generation Methods must learn strict amplitude, frequency, and phase information that determine a wireless signal's spectral and temporal identity, as well as the characteristics of the propagation channel through which it travels. This specific structure motivates physics-informed approaches for constructing and evaluating generative models for the wireless domain (Böck et al., 2025; Baur et al., 2025). A branch of prior work focuses on modeling the wireless channel, either by estimating it from received pilot measurements and known transmitted pilots using Wasserstein GANs (Balevi et al., 2020), score-based models (Arvinte & Tamir, 2022; 2023), or diffusion priors (Zhou et al., 2025), or by directly synthesizing channel realizations using location-conditioned diffusion models (Lee et al., 2026) or a VAE coupled to a parametric physics-based geometric channel model (Wagle et al., 2025). Generating I/Q sequences, however, is comparatively understudied and requires extra modeling for the joint variability of the transmitted waveform, propagation channel, and receiver noise over long sequences. We adapt RF-Diffusion, Time Weaver, and WaveStitch as baselines for this task (Chi et al., 2024; Narasimhan et al., 2024; Shankar et al., 2025).

3 Background

Consider a complex-valued baseband signal on the receiver side, $x(\tau) = I(\tau)+jQ(\tau)$, with discrete samples $x[n] = I[n]+jQ[n]$ for $n = 0, \ldots, N_{\text{samples}} - 1$. For N_{ant} receive antennas, stacking their synchronized sample sequences gives $\mathbf{X}_{\text{IQ}} \in \mathbb{C}^{N_{\text{ant}} \times N_{\text{samples}}}$, which we refer to as an IQ waveform. Each $\mathbf{X}_{\text{IQ}}$ is paired with heterogeneous conditioning information $\mathbf{Y}$, which may contain continuous or categorical quantities. After mapping each component to a fixed-dimensional real representation, such as a one-hot encoding for categorical variables, we write $\mathbf{Y} \in \mathbb{R}^{N_{\text{labels}}}$. A dataset is then $\mathcal{D} = \{(\mathbf{X}^i_{\text{IQ}}, \mathbf{Y}^i)\}_{i=1}^{N_{\text{pairs}}}$, where N_{pairs} is its size. For learning, we treat these pairs as independent and identically distributed (i.i.d.) samples from the empirical joint distribution $p_{\text{data}}(\mathbf{X}_{\text{IQ}}, \mathbf{Y})$.

Our primary objective is to learn a conditional generation model such that the IQ waveforms generated by the model with input label $\mathbf{Y}$ distributionally match the conditional distribution $p(\mathbf{X}_{\text{IQ}}|\mathbf{Y})$ for all valid $\mathbf{Y} \in \mathbb{R}^{N_{\text{labels}}}$ (Figure 1). To this end, we build on top of **Rectified Flow** (Liu et al., 2023b), which learns a velocity field that transports samples from a tractable source distribution to the target data distribution. Let $\mathbf{X}_0 \sim p_{\text{src}}$ denote a sample from a source distribution, such as $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and let $\mathbf{X}_1 \sim p_{\text{data}}$ denote a sample from the target data distribution. Rectified flow defines the straight-line interpolation, $\mathbf{X}_t = (1-t)\mathbf{X}_0 + t\mathbf{X}_1, t \in [0, 1]$, whose velocity, conditioned on the source-target pair $(\mathbf{X}_0, \mathbf{X}_1)$, is constant along the path $u_t = \frac{d\mathbf{X}_t}{dt} = \mathbf{X}_1 - \mathbf{X}_0$. We train a conditional time-dependent velocity field $v_\theta(\mathbf{X}_t, t, \mathbf{Y})$ using

$$\mathcal{L}_{\text{RF}} = \mathbb{E}_{\substack{(\mathbf{X}_1,\mathbf{Y})\sim p_{\text{data}},\, \mathbf{X}_0\sim p_{\text{src}},\\ t\sim\mathcal{U}[0,1]}} \left[\|v_\theta(\mathbf{X}_t, t, \mathbf{Y}) - (\mathbf{X}_1 - \mathbf{X}_0)\|_2^2 \right]. \tag{1}$$

Here, we uniformly sample the time t, *i.e.,* $t \sim \mathcal{U}[0, 1]$. Although each sampled pair defines a constant straight-line velocity, the regression objective learns the conditional mean transport induced by all such paths. At inference, we draw $\mathbf{X}_0 \sim p_{\text{src}}$ and integrate $d\mathbf{X}_t/dt = v_\theta(\mathbf{X}_t, t, \mathbf{Y})$ from $t = 0$ to $t = 1$ to obtain a sample from the learned conditional distribution.

4 WaveScene Dataset

WaveScene is a large-scale synthetic RF corpus of receiver-side baseband IQ waveforms generated through geometry-based stochastic channels and paired with respective labels. It covers 32 wave-

form classes and 1,066 parameterized variants spanning generic modulations, standardized communication protocols, and radar emissions. Each recording is a 4-antenna SIMO post-channel waveform captured by a horizontal uniform linear array with half-wavelength spacing ($N_{\text{ant}} = 4$), with $N_{\text{samples}} \in \{2048, 4096, 8192, 16384, 32768\}$. The corpus covers four propagation environments: urban microcell (UMi), urban macrocell (UMa), rural macrocell (RMa), and indoor hotspot (InH). Channel generation follows the geometry-based stochastic models specified in 3GPP TR 38.901 (3GPP, 2024b). The propagation states are line-of-sight (LOS), non-line-of-sight (NLOS), and outdoor-to-indoor (O2I), with frequencies determined by the scenario-specific channel models.

Table 1: Comparison with representative public IQ datasets.

Dataset	Coverage	Size	Length	Sampling rate	Multi-Rx	Varying f_c	Channel geometry
RadioML 2018 (O'Shea et al., 2018)	24 classes†	2.56M	1,024	N/A	×	×	×
Sig53 (Boegner et al., 2022)	53 classes†	6.51M	4,096	N/A	×	×	×
IQFM (Mashaal & Abou-Zeid, 2026)	7 classes†	1.12M	256	1, 10 MS/s	✓	×	×
EM-134K (Shen et al., 2026)	10 subsets‡	134K	1,024	20 MS/s	×	×	×
WaveScene	**32 classes**	**7.82M**	**2,048–32,768**	**96 kS/s–40 MS/s**	✓	✓	✓

† Modulation or simple waveform classes only; no protocol or radar waveform families. ‡ EM-134K is an instruction-tuning corpus spanning ten heterogeneous signal/task subsets, rather than mutually exclusive waveform classes.

The corpus contains approximately 7.82M recording-label pairs, including a 7.52M shared-channel set and a 300K channel-held-out set. Of this shared-channel set, 5.6M pairs form the development set, divided deterministically into 95% training and 5% validation data; the remaining 1.9M disjoint recordings form the seen-channel test set. We additionally construct the 300K channel-held-out test set whose channel realizations are unseen. Training uses 6 complementary labels. Waveform class (32), scene type (4), and propagation state (3) are categorical, while spectral occupancy, mean per-antenna pre-combining SNR, and RMS delay spread are continuous ($N_{\text{labels}} = 42$). Appendix C details the generation models and methods used.

5 Anchored Latent Flow

In this section, we introduce **Anchored Latent Flow** (ALF), a latent-space flow-matching model for efficient and coherent long-range receiver-side baseband IQ waveform generation conditioned on environment/propagation labels. As discussed in Section 2, we design ALF by treating the time-frequency inter-relationship in the RF-domain as a first-class inductive bias; see Appendix A for a formal justification. In particular, extending existing temporal generative models (Time Weaver, RF-Diffusion and WaveStitch) for long RF IQ sequences exposes two key challenges:

- **Poor Inference Efficiency.** Existing generators typically operate in the IQ space alone. For long-range IQ sequence generation, this results in high memory and inference costs. Particularly, this becomes evident for IQ waveforms of length 32,768 samples (Table 4). Moreover, 32k samples are still short; for example, a sequence of this length is roughly 10% of a single 20MHz LTE frame with a 30MHz sample rate.
- **Weak Long-Range Spectral Coherence.** The received IQ waveform evolution over time is globally constrained by spectral occupancy and the time-frequency structure. However, existing denoising or generative models do not explicitly constrain the model's representations to preserve this global structure. Consequently, the errors in IQ space can accumulate over time and produce globally inconsistent spectral behavior.

ALF addresses these two challenges through a two-stage training setup.

5.1 Stage 1: Label-Aligned Reconstruction-Aware Encoding

Rather than performing generation directly in IQ space, we operate in a compact *latent space representation*. More specifically, we learn a concise sequence of patch-level IQ latents that can be decoded back to the original waveform. For these patch-level latents, we use cross-patch attention to capture record-level context while preserving local waveform structure, and a SoftCLIP-style objective (Gao et al., 2024) to align a record-level summary with the conditioning labels.

Given a receiver-side baseband IQ waveform as described in Section 4, *i.e.,* $\mathbf{X}_{\text{IQ}} \in \mathbb{C}^{4 \times N_{\text{samples}}}$, we expose its real and imaginary components and divide it into $P = \lceil N_{\text{samples}}/N_{\text{patch}} \rceil$

non-overlapping patches, where N_{patch} is the patch length. Each patch is processed by a convolutional raw-IQ branch and a ButterflyNet-inspired (Li et al., 2020; Xu et al., 2020) spectral branch. A learned gate fuses the two representations, and after attention across the patches (global patch mixing), we get $\mathbf{Z}_{\text{IQ}} \in \mathbb{R}^{P\times 768}$. At $N_{\text{patch}} = 1024$, this corresponds to $\sim 11\times$ compression per patch. Concisely, we represent this as $\mathbf{Z}_{\text{IQ}} = \theta_{\text{enc-IQ}}(\mathbf{X}_{\text{IQ}})$. An IQ decoder parameterized by $\theta_{\text{dec-IQ}}$ mixes the latent sequence across patches and reconstructs the four-antenna waveform as $\widehat{\mathbf{X}}_{\text{IQ}} = \theta_{\text{dec-IQ}}(\mathbf{Z}_{\text{IQ}})$, while a `[CLS]` aggregate token further converts $\mathbf{Z}_{\text{IQ}}$ into a single record-level embedding $\mathbf{Z}_{\text{IQ-CLS}} \in \mathbb{R}^{768}$. Meanwhile, a label encoder θ_{lbl} maps the 6 labels of $\mathbf{X}_{\text{IQ}}$ to a corresponding 768-dimensional embedding $\mathbf{Z}_{\text{lbl}} = \theta_{\text{lbl}}(\mathbf{Y})$.

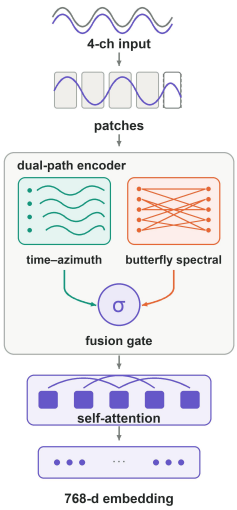


Figure 2: Fused dual-branch architecture with *ButterflyNet* (Li et al., 2020; Xu et al., 2020)

The contrastive objective therefore operates on the normalized IQ embedding $\mathbf{z}_i = \mathbf{Z}^i_{\text{IQ-CLS}}/\|\mathbf{Z}^i_{\text{IQ-CLS}}\|_2$ and the normalized label embedding $\mathbf{c}_j = \mathbf{Z}^j_{\text{lbl}}/\|\mathbf{Z}^j_{\text{lbl}}\|_2$. Their similarity is measured by the cosine logit $\mathbf{z}_i^\top \mathbf{c}_j/\tau$. The symmetric soft-target CLIP loss $\mathcal{L}_{\text{CLIP}}$ (Radford et al., 2021; Gao et al., 2024) assigns greater weight to pairs with matching categorical labels and nearby continuous attributes, allowing semantically similar records to act as partial positives rather than treating every non-paired record as an equally negative example. The reconstruction objective $\mathcal{L}_{\text{rec}}$, on the other hand, compares reconstructed $\widehat{\mathbf{X}}^i_{\text{IQ}}$ with $\mathbf{X}^i_{\text{IQ}}$ across complementary signal properties. Finally, a structural objective $\mathcal{L}_{\text{struct}}$ is used to avoid learned embedding from collapsing. Overall, we jointly train $\mathcal{L}_{\text{S1}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{CLIP}}(e)\mathcal{L}_{\text{CLIP}} + \lambda_{\text{struct}}(e)\mathcal{L}_{\text{struct}}$. During training, $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{struct}}$ are active throughout training. After an initial reconstruction-focused warm-up, $\mathcal{L}_{\text{CLIP}}$ is introduced, and $\lambda_{\text{CLIP}}(e)$, $\lambda_{\text{struct}}(e)$ are smoothly increased over subsequent epochs. Further implementation details and evaluations are provided in Appendix D.1. At the end of this stage, $\theta_{\text{enc-IQ}}$, $\theta_{\text{dec-IQ}}$, and θ_{lbl} are frozen, fixing the latent space for stage 2 (5.2).

5.2 Stage 2: Spectrally Anchored Latent Flow Matching

To explicitly incorporate RF time-frequency structure into generation, we introduce *spectral anchoring*. During flow matching (see Section 3), an intermediate representation of the velocity network is supervised to predict the corresponding spectrogram latent. Although the flow generates in the IQ latent space, our key insight is to constrain its intermediate representations, as in Figure 3. This encourages the generated waveform to follow the target's global spectral structure, such as occupied bandwidth, at sampled flow times. In Appendix A, we present a stylized model that shows spectral anchoring exposes the low-dimensional structure induced by a bandwidth constraint, reducing the parameter complexity of the feasible signal distribution relative to an unstructured time-domain parameterization.

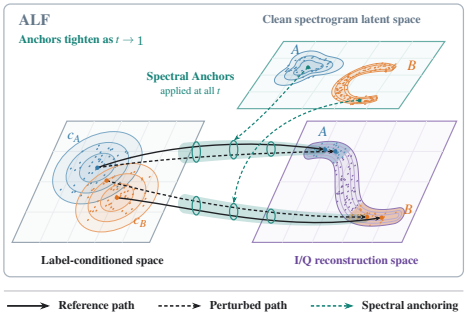


Figure 3: Spectral anchoring guides learned flow with frequency-domain latents.

Let $\mathbf{Z}_1$ denote the standardized IQ latent corresponding to the IQ waveform $\mathbf{X}_{\text{IQ}}$. Since our objective is to learn the label-conditioned data distribution $p_{\text{data}}(\cdot \mid \mathbf{Y})$, rather than initializing the flow from the same standard normal distribution for every label, we use the frozen probabilistic label encoder from Stage 1, $(\boldsymbol{\mu}_{\text{lbl}}(\mathbf{Y}), \log \boldsymbol{\sigma}^2_{\text{lbl}}(\mathbf{Y})) = \theta_{\text{lbl}}(\mathbf{Y})$, to define the label-conditioned source distribution (implementation in E.2). For a recording containing P patches, we independently sample $\mathbf{Z}_0^{(p)} \sim \mathcal{N}\big(\boldsymbol{\mu}_{\text{lbl}}(\mathbf{Y}), \text{diag}(\boldsymbol{\sigma}^2_{\text{lbl}}(\mathbf{Y}))\big)$ for $p = 1, \ldots, P$. These samples form $\mathbf{Z}_0$, and the conditional velocity network $v_\theta^{\text{cond}}(\mathbf{Z}_t, t, \mathbf{Y})$ learns to transport them to $\mathbf{Z}_1$.

Importantly, to anchor the flow with frequency-domain information, we separately train a spectral autoencoder on the level-carrying log-power spectrogram associated with each IQ waveform (implementation in E.3). This autoencoder is distinct from the ButterflyNet-inspired branch of the Stage-1 IQ encoder. Its post-mixing latent sequence, denoted $\mathbf{Z}_{\text{spec}}$, retains both patch-level spectral content and recording-level context. We then freeze the spectral autoencoder and independently standardize its output to obtain $\widetilde{\mathbf{Z}}_{\text{spec}}$, which serves as the target for the anchoring objective.

To connect this spectral target to the flow, we attach a lightweight prediction head θ_{anchor} to the hidden representation after the sixth Transformer block (Figure 1). At each sampled flow time t, the head maps $\mathbf{H}_t^{(6)}$ to an estimate of the standardized clean spectrogram latent $\widetilde{\mathbf{Z}}_{\text{spec}}$. The target remains fixed across flow times, while its loss contribution is weighted by $t/\bar{t}$ to emphasize states closer to the target IQ distribution. We sample $t \sim p_t$ by drawing $\epsilon_t \sim \mathcal{N}(0,1)$ and setting $t = \text{sigmoid}(\epsilon_t)$, with $\bar{t} = \mathbb{E}_{t\sim p_t}[t]$. The anchor head and velocity network are trained jointly with the following loss:

$$\mathcal{L}_{\text{ALF}} = \mathbb{E}_{(\mathbf{Z}_1,\mathbf{Y},\widetilde{\mathbf{Z}}_{\text{spec}})\sim p_{\text{data}};\, \mathbf{Z}_0\sim p_{\text{src}}(\cdot|\mathbf{Y});\, t\sim p_t} \left[\underbrace{\left\| v_\theta^{\text{cond}}(\mathbf{Z}_t, t, \mathbf{Y}) - (\mathbf{Z}_1 - \mathbf{Z}_0) \right\|_2^2}_{\text{Flow-matching loss}} + \underbrace{\lambda_{\text{anchor}}(e)\, \frac{t}{\bar{t}}\, \left\| \theta_{\text{anchor}}\left(\mathbf{H}_t^{(6)}\right) - \widetilde{\mathbf{Z}}_{\text{spec}} \right\|_2^2}_{\text{Spectral anchoring loss}} \right]. \quad (2)$$

We ramp $\lambda_{\text{anchor}}(e)$ to 0.25 over the first two training epochs. The first term in Equation 2 learns the label-conditioned transport from $\mathbf{Z}_0$ to $\mathbf{Z}_1$, while the second shapes the sixth-block representation using the target spectrogram latent. At inference, the generated $\mathbf{Z}_1$ is reconstructed using the frozen decoder $\theta_{\text{dec-IQ}}$; the spectral autoencoder and anchor head are omitted.

6 RESULTS

Table 2: TSTR/TRTS balanced accuracies for downstream objectives, by length (k = num classes)

		Waveform (k=32)		Bandwidth (k=9)		Sample Rate (k=13)		Carrier Freq (k=30)	
Length	Method	TSTR	TRTS	TSTR	TRTS	TSTR	TRTS	TSTR	TRTS
2,048	WaveStitch	0.02	0.04	0.11	0.11	0.08	0.09	0.04	0.04
	RF Diffusion†	0.03	0.03	0.11	0.11	0.08	0.08	0.04	0.04
	Time Weaver	0.59	0.33	0.39	0.26	0.39	0.21	0.24	0.16
	ALF (ours)	**0.84**	**0.83**	**0.63**	**0.56**	**0.70**	**0.68**	**0.27**	**0.30**
4,096	WaveStitch	0.03	0.03	0.12	0.11	0.07	0.08	0.03	0.03
	Time Weaver	0.60	0.37	0.44	0.29	0.41	0.17	0.24	0.16
	ALF (ours)	**0.85**	**0.82**	**0.62**	**0.58**	**0.67**	**0.65**	**0.29**	**0.32**
8,192	WaveStitch	0.03	0.03	0.12	0.11	0.08	0.08	0.03	0.04
	Time Weaver	0.61	0.39	0.46	0.29	0.42	0.19	0.23	0.18
	ALF (ours)	**0.84**	**0.81**	**0.61**	**0.57**	**0.70**	**0.66**	**0.28**	**0.30**
16,384	WaveStitch	0.03	0.03	0.12	0.11	0.07	0.08	0.03	0.03
	Time Weaver	0.60	0.43	0.45	0.31	0.43	0.20	0.22	0.19
	ALF (ours)	**0.83**	**0.81**	**0.61**	**0.58**	**0.70**	**0.66**	**0.28**	**0.29**
32,768	WaveStitch	0.03	0.03	0.12	0.11	0.07	0.08	0.03	0.03
	Time Weaver	0.62	0.44	0.52	0.37	0.44	0.24	0.23	0.20
	ALF (ours)	**0.83**	**0.79**	**0.70**	**0.64**	**0.72**	**0.68**	**0.28**	**0.30**
Avg.	WaveStitch	0.03	0.03	0.12	0.11	0.07	0.08	0.03	0.03
	Time Weaver	0.61	0.40	0.46	0.31	0.42	0.20	0.23	0.18
	ALF (ours)	**0.84**	**0.81**	**0.64**	**0.59**	**0.70**	**0.66**	**0.28**	**0.30**
ALF Advantage	**vs. Time Weaver**	**38.02%**	**102.88%**	**38.29%**	**91.51%**	**64.09%**	**228.97%**	**21.47%**	**66.69%**

† *RF Diffusion could not be trained for recordings of length greater than 2,048 due to memory limitations; longer recordings were not attempted.*

Due to the highly structured nature of RF signals, the differences between valid and undecipherable sequences are hard to differentiate using standard metrics. Instead, it has been noted in the wireless community that demonstrating downstream performance is a better judge of generation quality (Baur et al., 2025). To this end, we can show utility to waveform, channel, and receiver classification using our chosen labels (Appendix C.4). Although distributional metrics are insufficient measures

of quality in this domain, comparisons across methods using PSD and Spectrogram MSEs can be found in Appendix G.1. Table 2 compares ALF against related time series methods, all trained on the same 6 label conditions.

Procedures: Classifications use lightweight, length-agnostic, 1-D CNN classifiers trained for 8,000 steps across generated corpora of 100k recordings and evaluated on held-out test sets of size 35,856. Permuting generated (synthetic) and reference (real) data as training and evaluation sets produces various metrics that characterize a model's capability in a downstream classification job.

Baselines: We compare conditional time-series generative models from Narasimhan et al. (2024), Chi et al. (2024), and Shankar et al. (2025). These were selected for their RF focus or state-of-the-art metadata-conditioned generation. Table 2 shows how WaveStitch and RF Diffusion experience a total collapse when evaluated on downstream tasks. Time Weaver and ALF are the only models capable of achieving any downstream utility (Narasimhan et al., 2024).

Metrics: Train-on-Synthetic, Test-on-Real **(TSTR)** attempts to classify unseen, *real* data after learning to recognize generated data, alluding to a corpus' indistinguishability from the truth. This requires a diverse range of representative generations that span all "tastes" of a signal—corpus generality. Where TSTR examines the entire dataset, Train-on-Real, Test-on-Synthetic **(TRTS)** evaluates if a synthetic dataset can be categorized correctly using an understanding of real signal features. These accuracies serve as certificates of individual recording-level fidelity.

ALF shows high utility over baselines on frequency-dependent tasks. With respect to waveform classification, ALF shows significant improvements of 38.02% TSTR and 102.88% TRTS performances (Table 2). TSTR confirms that generations have generality, and TRTS confirms recording-level specificity. Bandwidth and sample rate classification improvements demonstrate how *spectrogram reasoning* learns representations of derived labels (Appendix C.4). Furthermore, utility is shown for unseen labels in carrier frequency estimation. Compared to other models, ALF exploits the domain-specific biases necessary for downstream evaluations. More seen and unseen classifications (Indoor/Outdoor, Carrier Frequency, Delay Spread) can be found in Appendix G.2.

Table 3: (a) Anchored vs. unanchored ablations across RF metrics. (b) Reclassification under low-density samples $\mathcal{N}(\mu, (k\sigma)^2)$. 50 steps per generation.

(a) Ablation	Performance				Gain vs. base LF (%)			
Method	TSTR† ↑	PSD ↓	PAPR ↓	CV ↓	TSTR	PSD	PAPR	CV
LF-N	81.93	17.87	4.99	0.139	—	—	—	—
LF-M	83.49	17.48	4.87	0.133	+1.9	+2.2	+2.4	+3.9
ALF	**84.49**	**16.86**	**4.70**	**0.126**	**+3.1**	**+5.6**	**+5.7**	**+9.3**

(b) Robustness	Accuracy (%) ↑			ALF gain over X (%)		
Method X	$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$
ALF	**96.5**	**53.6**	**11.7**	—	—	—
LF-M	92.6	45.0	7.8	**+4.2**	**+19.2**	**+50.0**
LF-N	90.1	12.5	3.9	**+7.1**	**+329.5**	**+200.0**
Time Weaver	39.0	4.6	3.3	**+147.3**	**+1075.3**	**+257.1**

Note: LF-M/LF-N: unanchored latent flow with metadata/standard-normal priors; (a) †Seed 100. PSD: median PSD MSE (dB^2); PAPR: error (dB); CV: envelope-CV error. (b) classifiers pretrained on ground truth. All ablations use a refiner for SNR over 30db (Appendix E.6)

Anchoring improves RF-specific metrics, sampling robustness, and parameter complexity. Before presenting numerical ablations, we note the following. With an "informative" (support-preserving) spectral cross-domain anchor, we theoretically show that bandwidth constraints in the spectrum reduces model parameter complexity (in effect, a smaller model suffices for high-fidelity generation). A detailed exposition is in Appendix A.

Next, Table 3 (a) shows that spectral anchoring provides gains beyond metadata priors alone for TSTR and specific RF amplitude properties. Together, PSD, PAPR, and envelope CV assess spectral power distribution, relative peak power, and overall amplitude variability. ALF outperforms both label-source initialization and base LF, with consistent gains in spectral and I/Q amplitude fidelity. Table 3 (b) demonstrates how anchoring slows the rate of generation collapse when sampling from low-density areas ($k\sigma$ from mean μ).

Anchoring reasons through spectrotemporal structure. To inspect the impact of latent space anchoring, we compare decoded intermediate representations of ALF generation trajectories to an unanchored latent flow (LF) ablation using the pre-trained spectrogram autoencoder. Figure 4 shows that ALF reasons an identifiable Costas hopping pattern, whereas the baseline shape remains ambiguous. As lengths scale, sharper frequency patterns better constrain time-domain generation, resulting in compounding sample coherence advantages.

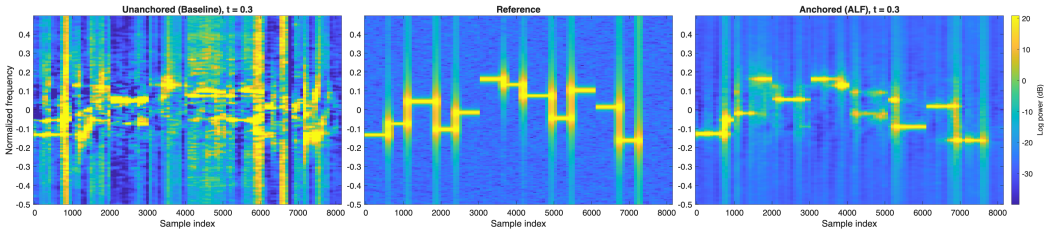


Figure 4: Visualization of ALF vs. LF (unanchored) latent space as decoded spectrograms. LF borrows the lightweight projection head from ALF.

ALF reduces the inference bottleneck. This is a major barrier to meaningful wireless signal generation at-length: it is time-prohibitive to generate long sequences. Although inference complexity remains bound by attention scaling laws (Vaswani et al., 2017), ALF's latent reasoning architecture and per-patch decoding manifest as two orders of magnitude in inference time reduction to Time Weaver, along with its generative quality boost (see Tables 4 and 2). Other models shorten this gap, but at the cost of unusable data. These gains make lengths of 100k+ feasible in fractions of a second.

Table 4: Latency and allocated-memory comparison at B=16 on one NVIDIA H100 GPU.

Batch	$N_{samples}$	Latency (ms/rec)			Allocated Memory (MiB)		
		ALF	TimeWeaver	ALF vs. TimeWeaver	**ALF**	TimeWeaver	ALF vs. TimeWeaver
16	2048	**25.5**	447.1	18×	1045.1	**816.0**	0.8×
16	4096	**25.9**	833.1	32×	**1173.4**	1272.0	1.1×
16	8192	**25.8**	1718.0	67×	**1431.9**	2184.0	1.5×
16	16384	**25.7**	3514.1	137×	**1947.1**	4006.0	2.1×
16	32768	**34.4**	7242.1	211×	**2976.8**	7652.0	2.6×
16	**Avg**	**27.5**	**2750.9**	**100×**	**1714.9**	**3186.0**	**1.9×**

Our two-stage architecture enables native and efficient style transfer. The field ALF learns flows from label-conditioned source latents $\mathbf{Z}_0$ to signal latents $\mathbf{Z}_1$ (§ 5.2). The learned field can also be integrated backward from $t = 1$ to $t = 0$, without a stochastic reverse process. Thus, ALF has a clean analysis/synthesis pipeline, making it a model native to style transfer. We edit held-out recordings by encoding each input signal and reversing the learned field to obtain its source-space representation. Using a Gaussian log-likelihood search, we find the most likely label for that representation.

Style transfer, including signal denoising (see Figure 5), is executed in one of two ways. **Fresh Prior Resample**: Given the estimated label $\hat{c}$, increase SNR and regenerate by sampling from the higher SNR mode. **RF-Inversion**: Given the observed record and a higher SNR target, use a controller at inference to guide it to the corrected region of the reconstruction space (Rout et al., 2025b). Figure 5 showcases ALF's capability of denoising recordings that are near or below the noise floor; likewise, this demonstrates learned signal representations in harsher SNR regimes. Notably,

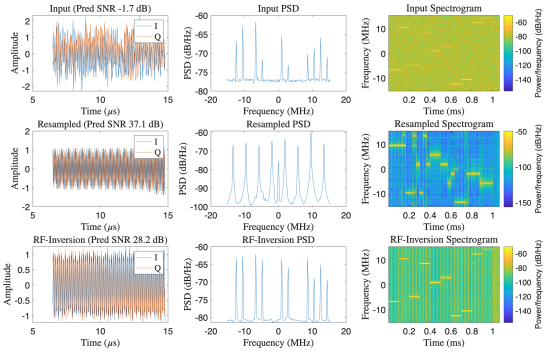


Figure 5: Resampling vs. RF-Inversion style transfer

a frequency-hopped frequency modulation (FHFM) example shows the difference in output similarity between naive resampling and a controlled approach (Rout et al., 2025b). Notably, although RF-Inversion marginally decreases the generated SNR, this approach maintains the original signal information in the hopping sequence while still effectively denoising. We also show how latent anchor guidance helps prevent occasional spectral blurring (E.8).

Autoregressive generation unlocks arbitrary-length sequences and distribution shifts. We demonstrate this adaptation of ALF through *suffix generation*. This post-training learns to flow new patches alongside fixed previously generated tokens. Thus, new sections now learn two conditions: their target label, and a context window of the signal. Once in the reconstruction space, patches are decoded without their prefix tokens. We provide additional information in Appendix F.

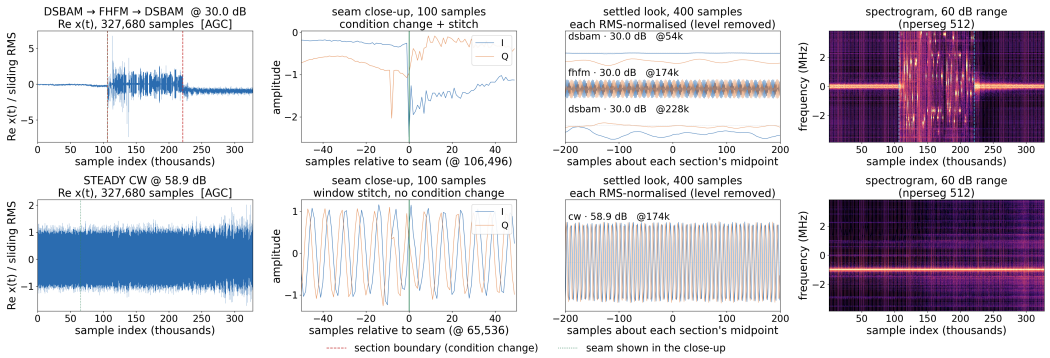


Figure 6: 327k-sample-long AR generations show waveform shifts and single-condition continuity

Limitations: It is notable that ALF does not show improvement on scene- or propagation-centered tasks. These binary classifications are characterized by energy fluctuations which are normalized away, resulting in noise readings (§ G, Table 15). However, TRTS metrics show ALF has equal, if not better, features across individual time series. This suggests that although our model's corpus is less general than Time Weaver in these tasks, records maintain scene-specific features.

Our results suggest that spectral anchoring shapes the flow's intermediate representations and improves generation without adding inference-time computation. Nevertheless, several limitations remain. First, WaveScene has limited payload diversity and uses simulated channels that remain static within each recording, omitting mobility and Doppler evolution, hardware impairments, interference, and other variability present in real transmissions. Second, the current anchor is derived from a log-power spectrogram and therefore emphasizes time-frequency energy without explicitly encoding spectral phase or cross-antenna relationships. Finally, the present evaluation focuses primarily on distributional similarity and downstream classification; broader evaluation of reversible-flow transformations, receiver-level performance, and over-the-air recordings remains necessary.

7 Conclusion

We introduced ALF, a latent flow-matching framework for conditional generation of long, multi-antenna IQ recordings, together with WaveScene, a dataset spanning diverse waveform classes, propagation conditions, and record lengths. By combining metadata-conditioned initialization with training-time spectral anchoring, ALF improves spectral fidelity and downstream utility without adding inference-time computation. Across held-out channels and all evaluated lengths, it reaches an average TSTR accuracy of 84.5%, a 38.02% relative improvement over SOTA conditional time-series generator Time Weaver while generating records $100\times$ faster on average.

Future work. Important next steps include training and evaluating on over-the-air recordings, including receiver-level validation under hardware and environmental impairments. Alternative flow constraints and anchor representations could incorporate phase, cross-antenna structure, multiple time-frequency resolutions, or supervision at different flow layers and timesteps. Extending the model to substantially longer recordings and stateful streaming generation would further test whether the architecture can maintain coherence beyond the sequence lengths considered here.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in implementing our method and got feedback for the method/experiments. We also used generative AI tools to assist in the writing and checking of proofs in the Appendix (verifying derivations, identifying and correcting errors in intermediate formulas, and drafting revised versions of appendix proofs), to provide feedback on the research and its presentation. We did not use it to build our theoretical/conceptual framework, refining our hypothesis, generating synthetic data, support qualitative and thematic data analysis and interpret results. Additionally, we used AI tools to edit/polish our written draft, edit existing codebases, summarize existing literature, identifying literature relevant to our method. We have reviewed the code generated by AI tools and take responsibility for the final content of this work, including the text, claims or artifacts produced with the aid of generative AI.

REFERENCES

3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation. Technical Specification TS 36.211, 3rd Generation Partnership Project, April 2023. URL https://www.etsi.org/deliver/etsi_ts/136200_136299/136211/17.03.00_60/ts_136211v170300p.pdf. Release 17.

3GPP. NR; Physical Channels and Modulation. Technical Specification TS 38.211, 3rd Generation Partnership Project, October 2024a. URL https://www.etsi.org/deliver/etsi_ts/138200_138299/138211/18.04.00_60/ts_138211v180400p.pdf. Release 18.

3GPP. Study on channel model for frequencies from 0.5 to 100 GHz. Technical Report TR 38.901, 3rd Generation Partnership Project, May 2024b. URL https://www.etsi.org/deliver/etsi_tr/138900_138999/138901/18.00.00_60/tr_138901v180000p.pdf. Version 18.0.0, Release 18.

Ahmed Aboulfotouh and Hatem Abou-Zeid. Multimodal wireless foundation models. *arXiv preprint arXiv:2511.15162*, 2025. URL https://arxiv.org/abs/2511.15162.

Ahmed Aboulfotouh, Elsayed Mohammed, and Hatem Abou-Zeid. 6g wavesfm: A foundation model for sensing, communication, and localization. *IEEE Open Journal of the Communications Society*, 6:6792–6807, 2025. doi: 10.1109/OJCOMS.2025.3600616.

Advanced Television Systems Committee. ATSC Digital Television Standard, Part 2: RF/Transmission System Characteristics. Technical Report A/53 Part 2:2011, Advanced Television Systems Committee, December 2011. URL https://www.atsc.org/wp-content/uploads/2015/03/a_53-Part-2-2011.pdf.

Juan Lopez Alcaraz and Nils Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models. *Transactions on Machine Learning Research*, 2023a. ISSN 2835-8856. URL https://openreview.net/forum?id=hHiIbk7ApW.

Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based conditional ecg generation with structured state space models. *Computers in Biology and Medicine*, 163:107115, 2023b. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2023.107115. URL https://www.sciencedirect.com/science/article/pii/S0010482523005802.

Marius Arvinte and Jonathan I. Tamir. Score-based generative models for robust channel estimation. In *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 453–458, 2022. doi: 10.1109/WCNC51071.2022.9771907.

Marius Arvinte and Jonathan I. Tamir. Mimo channel estimation using score-based generative models. *IEEE Transactions on Wireless Communications*, 22(6):3698–3713, 2023. doi: 10.1109/TWC.2022.3220784.

Eren Balevi, Akash Doshi, Ajil Jalal, Alexandros Dimakis, and Jeffrey G. Andrews. High dimensional channel estimation using deep generative networks, 2020. URL https://arxiv.org/abs/2006.13494.

Michael Baur, Nurettin Turan, Simon Wallner, and Wolfgang Utschick. Evaluation metrics and methods for generative models in the wireless phy layer. *IEEE Transactions on Machine Learning in Communications and Networking*, 3:677–689, 2025. doi: 10.1109/TMLCN.2025.3571026.

Marin Biloš, Kashif Rasul, Anderson Schneider, Yuriy Nevmyvaka, and Stephan Günnemann. Modeling temporal data as continuous functions with stochastic process diffusion. In *international conference on machine learning*, pp. 2452–2470. PMLR, 2023.

Bluetooth SIG. Bluetooth Core Specification, Version 5.4, January 2023. URL https://www.bluetooth.com/wp-content/uploads/Files/Specification/HTML/Core-54/out/en/index-en.html.

Benedikt Böck, Andreas Oeldemann, Timo Mayer, Francesco Rossetto, and Wolfgang Utschick. Physics-informed generative modeling of wireless channels. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 4602–4626. PMLR, 2025. URL https://proceedings.mlr.press/v267/bock25a.html.

Luke Boegner, Manbir Gulati, Garrett Vanhoy, Phillip Vallance, Bradley Comar, Silvija Kokaljl-Filipovic, Craig Lennon, and Robert D Miller. Large scale radio frequency signal classification. *arXiv preprint arXiv:2207.09918*, 2022.

Guoxuan Chi, Zheng Yang, Chenshu Wu, Jingao Xu, Yuchong Gao, Yunhao Liu, and Tony Xiao Han. RF-Diffusion: Radio signal generation via time-frequency diffusion. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, ACM MobiCom '24, pp. 77–92, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704895. doi: 10.1145/3636534.3649348. URL https://doi.org/10.1145/3636534.3649348.

DiffusionGemma Team, Adrien Ali Taïga, et al. Diffusiongemma technical report. *arXiv preprint arXiv:2608.00146*, 2026.

ETSI. Digital Video Broadcasting (DVB); Second Generation Framing Structure, Channel Coding and Modulation Systems for Broadcasting, Interactive Services, News Gathering and Other Broadband Satellite Applications; Part 1: DVB-S2. EN 302 307-1 V1.4.1, European Telecommunications Standards Institute, November 2014. URL https://www.etsi.org/deliver/etsi_en/302300_302399/30230701/01.04.01_60/en_30230701v010401p.pdf.

Federal Communications Commission. Table of Frequency Allocations. 47 C.F.R. § 2.106, 2026. URL https://www.ecfr.gov/current/title-47/chapter-I/subchapter-A/part-2/subpart-B/section-2.106. Accessed: 2026-09-14.

Yuting Gao, Jianjie Jiao, Jinfeng Zhang, Weichen Wang, Shuyan He, Lin Shen Shen, Ke Li, and Wei-Shi Zheng. Softclip: Softer cross-modal alignment makes clip stronger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1860–1868, 2024. doi: 10.1609/aaai.v38i3.27955.

Albert Gu, Karan Goel, and Christopher Re. Efficiently Modeling Long Sequences with Structured State Spaces. In *International Conference on Learning Representations*, 2022. URL https://mlanthology.org/iclr/2022/gu2022iclr-efficiently/.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/39235c56aef13fb05a6adc95eb9d8d66-Abstract.html.

Jakob Hoydis, Sebastian Cammerer, Fayçal Ait Aoudia, Avinash Vem, Nikolaus Binder, Guillermo Marcus, and Alexander Keller. Sionna: An open-source library for next-generation physical layer research, 2023. URL https://arxiv.org/abs/2203.11854.

IEEE. IEEE Standard for Information Technology—Telecommunications and Information Exchange Between Systems Local and Metropolitan Area Networks—Specific Requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, 2020a. URL https://standards.ieee.org/ieee/802.11/7028/.

IEEE. IEEE Standard for Low-Rate Wireless Networks, 2020b. URL https://standards.ieee.org/standard/802_15_4-2020.html.

IEEE. IEEE Standard for Information Technology—Telecommunications and Information Exchange Between Systems Local and Metropolitan Area Networks—Specific Requirements—Part 11: Wireless LAN MAC and PHY Specifications Amendment 1: Enhancements for High-Efficiency WLAN, 2021. URL https://standards.ieee.org/ieee/802.11ax/7180/.

Stephan Jaeckel, Leszek Raschkowski, Kai Börner, and Lars Thiele. Quadriga: A 3-d multi-cell channel model with time evolution for enabling virtual field trials. *IEEE Transactions on Antennas and Propagation*, 62(6):3242–3256, 2014. doi: 10.1109/TAP.2014.2310220.

Marcel Kollovieh, Abdul Fatir Ansari, Michael Bohlke-Schneider, Jasper Zschiegner, Hao Wang, and Yuyang Bernie Wang. Predict, refine, synthesize: Self-guiding diffusion models for probabilistic time series forecasting. *Advances in Neural Information Processing Systems*, 36:28341–28364, 2023.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=a-xFK8Ymz5J.

Taekyun Lee, Juseong Park, Hyeji Kim, and Jeffrey G. Andrews. Generating high dimensional user-specific wireless channels using diffusion models. *IEEE Transactions on Wireless Communications*, 25:2907–2921, 2026. doi: 10.1109/TWC.2025.3600286.

Nadav Levanon and Eli Mozeson. *Radar Signals*. Wiley, 2004. doi: 10.1002/0471711219.

Yan Li, Xinjiang Lu, Yaqing Wang, and Dejing Dou. Generative time series forecasting with diffusion, denoise, and disentanglement. *Advances in Neural Information Processing Systems*, 35: 23009–23022, 2022.

Yingzhou Li, Xiuyuan Cheng, and Jianfeng Lu. Butterfly-net: Optimal function representation based on convolutional neural networks. *Communications in Computational Physics*, 28(5):1838–1885, 2020. doi: 10.4208/cicp.OA-2020-0214. URL https://doi.org/10.4208/cicp.OA-2020-0214.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/liu23f.html.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=XqjTW_wxZ9Y.

Omar Mashaal and Hatem Abou-Zeid. IQFM—a wireless foundation model for I/Q streams in ai-native 6g. *IEEE Open Journal of the Communications Society*, 7:1426–1441, 2026. doi: 10.1109/OJCOMS.2026.3661435. URL https://arxiv.org/abs/2506.06718.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

Sai Shankar Narasimhan, Shubhankar Agarwal, Oguzhan Akcin, Sujay Sanghavi, and Sandeep P. Chinchali. Time weaver: A conditional time series generation model. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=WpKDeixmFr.

Sai Shankar Narasimhan, Shubhankar Agarwal, Litu Rout, Sanjay Shakkottai, and Sandeep P. Chinchali. Constrained posterior sampling: Time series generation with hard constraints. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=iBLAYaEKdL.

Timothy James O'Shea, Tamoghna Roy, and T. Charles Clancy. Over-the-air deep learning based radio signal classification. *IEEE Journal of Selected Topics in Signal Processing*, 12(1):168–179, 2018. doi: 10.1109/JSTSP.2018.2797022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

John G. Proakis and Masoud Salehi. *Digital Communications*. McGraw-Hill, 5 edition, 2008.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8857–8868. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/rasul21a.html.

Mark A. Richards, James A. Scheer, and William A. Holm. *Principles of Modern Radar: Basic Principles*. SciTech Publishing, 2010.

Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Anchored diffusion language model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=E8adS5srds.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=Hu0FSOSEyS.

Semtech Corporation. SX1276/77/78/79: 137 MHz to 1020 MHz Low Power Long Range Transceiver, 2020. URL https://www.semtech.com/products/wireless-rf/lora-connect/sx1276.

Aditya Shankar, Lydia Chen, Arie van Deursen, and Rihan Hai. Wavestitch: Flexible and fast conditional time series generation with diffusion models. *Proceedings of the ACM on Management of Data*, 3(6 (SIGMOD)):1–25, dec 2025. doi: 10.1145/3769842. URL https://doi.org/10.1145/3769842.

Mingyuan Shao, Dingzhao Li, Shaohua Hong, Jie Qi, and Haixin Sun. IQFormer: A novel transformer-based model with multi-modality fusion for automatic modulation recognition. *IEEE Transactions on Cognitive Communications and Networking*, 11(3):1623–1634, 2025. doi: 10.1109/TCCN.2024.3485118.

Junyu Shen, Zhendong She, Chenghanyu Zhang, Yuchuang Sun, Luqing Luo, Dingwei Tan, Zonghao Guo, Bo Guo, Zehua Han, Wupeng Xie, Yaxin Mu, Peng Zhang, Peipei Li, Fengxiang Wang, Yangang Sun, and Maosong Sun. Merlin: Building low-snr robust multimodal llms for electromagnetic signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8653–8663, June 2026.

Merrill I. Skolnik (ed.). *Radar Handbook*. McGraw-Hill, 3 edition, 2008. ISBN 978-0071485470.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265. PMLR, 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.02502.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24804–24816, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cfe8504bda37b575c70ee1a8276f3486-Abstract.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/7181-attention-is-all-you-need.

Satyavrat Wagle, Akshay Malhotra, Shahab Hamidi-Rad, Aditya Sant, David J. Love, and Christopher G. Brinton. Physics-informed generative approaches for wireless channel modeling, 2025. URL https://arxiv.org/abs/2503.05988.

P. Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967. doi: 10.1109/TAU.1967.1161901.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Zhongshu Xu, Yingzhou Li, and Xiuyuan Cheng. Butterfly-Net2: Simplified Butterfly-Net and Fourier transform initialization. In *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pp. 431–450. PMLR, 2020. URL https://proceedings.mlr.press/v107/xu20b.html.

Tijin Yan, Hongwei Zhang, Tong Zhou, Yufeng Zhan, and Yuanqing Xia. Scoregrad: Multivariate probabilistic time series forecasting with continuous energy-based generative models. *arXiv preprint arXiv:2106.10121*, 2021.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*, 2025.

Xinyu Yuan and Yan Qiao. Diffusion-ts: Interpretable diffusion for general time series generation. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 41582–41610, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/b5b66077d016c037576cc56a82f97f66-Paper-Conference.pdf.

Xingyu Zhou, Le Liang, Jing Zhang, Peiwen Jiang, Yong Li, and Shi Jin. Generative diffusion models for high dimensional channel estimation. *IEEE Transactions on Wireless Communications*, 24 (7):5840–5854, 2025. doi: 10.1109/TWC.2025.3549592.

APPENDIX

A SPECTRAL ANCHORING PARAMETER COMPLEXITY

A.1 OVERVIEW

In this section, we argue that spectral anchoring is particularly effective for wireless signal generation. Specifically, anchoring by the spectrogram explicitly exposes the low-dimensional structure induced by band limit constraints, thereby reducing the parameter complexity of an unstructured time-domain representation. Consider the signal,

$$x = (x_0, \ldots, x_{n-1}) \in \mathcal{D}^n.$$

Generally, such a signal has n degrees of freedom. Let $d = |\mathcal{D}|$. Since probabilities must sum to 1, there are $d^n - 1$ free parameters[1]. If the signal is band-limited to k frequency bins, however, the DFT of x exposes a useful structure (see below in Appendix A.2 for formal definitions).

$$\hat{x} = (\underbrace{0, \ldots, 0}_{\text{known}}, \hat{x}_{j_1}, \ldots, \hat{x}_{j_k}, \underbrace{0, \ldots, 0}_{\text{known}}).$$

We can observe that the signal has k degrees of freedom under the Fourier basis. The key insight is that, because the Fourier transform is invertible, this reduction in effective degrees of freedom also constrains the time-domain signal. While band limitation is explicit in the frequency domain through the restriction to k active coefficients, the corresponding structure is implicit in the time domain, where it manifests as dependencies across the n samples. Spectral anchoring reveals these constraints directly to the generative model, reducing the parameter complexity required relative to an unstructured time-domain parameterization. In the subsequent exposition, using a stylized simple model, we formalize this intuition. While this model differs from the details of our implementation, we believe that it captures some of the core intuition motivating our approach.

A.2 SETUP

Signals and spectra. Let $\mathcal{D} \subset \mathbb{C}$ be a finite set of $d = |\mathcal{D}|$ possible I/Q values, and let

$$x = (x_0, \ldots, x_{n-1}) \in \mathcal{D}^n$$

denote a length-n I/Q sequence, where each sample x_i takes a value in $\mathcal{D}$.

The DFT of x is

$$\hat{x}_j = \sum_{t=0}^{n-1} x_t e^{-2\pi i j t/n}, \qquad j = 0, \ldots, n-1. \tag{3}$$

Now we will define a set of band-limited signals. For a set of permissible frequency indices $K \subseteq \{0, \ldots, n-1\}$, let

$$\mathcal{X}_K = \left\{ x \in \mathcal{D}^n : \hat{x}_j = 0 \text{ for all } j \notin K \right\} \tag{4}$$

be the signals whose spectrum is supported on K.

[1] A generative model, in essence, learns a joint distribution. With discrete/finite-alphabet, equivalently, it learns to fill-in the entries of a joint/conditional distribution table (as example, see the table for $p_{Y|C}(y \mid c)$ in Appendix A.2 below), where each entry in this table can be thought of as a "parameter" that is to be learned. Dependencies across entries in the table can be exploited to reduce the number of parameters to be learned, leading to the "free parameter" count. A smaller number of free parameters implies that a smaller capacity model suffices to learn a distribution with high fidelity. The free parameter count can be translated to a training sample complexity bound using standard concentration inequalities. Concretely, suppose that an ϵ-accurate estimate is needed for each free parameter entry, with probability at least $1 - \delta$. A standard concentration bound (Hoeffding/Chernoff) implies that each entry requires order of $\epsilon^{-2} \ln(1/\delta)$ samples. More generally, learning a distribution with D free parameters (i.e., supported on $D + 1$ outcomes) to within total variation distance ϵ requires order of $(D + \ln(1/\delta))/\epsilon^2$ samples, so the sample complexity grows linearly with the free parameter count. Thus in this section, we focus directly on the free parameter analysis for spectral anchoring.

A.3 Unanchored Parameter Count

Counting Strategy. To count the free parameters required to represent a conditional distribution $p_{Y|C}(y \mid c)$, suppose that C can take L possible values $c_1, \ldots, c_L$ and that Y can take d possible values $y_1, \ldots, y_d$. We represent the distribution as a conditional probability table,

	$Y = y_1$	$Y = y_2$	$\cdots$	$Y = y_d$
$C = c_1$	$p(y_1 \mid c_1)$	$p(y_2 \mid c_1)$	$\cdots$	$p(y_d \mid c_1)$
$C = c_2$	$p(y_1 \mid c_2)$	$p(y_2 \mid c_2)$	$\cdots$	$p(y_d \mid c_2)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$C = c_L$	$p(y_1 \mid c_L)$	$p(y_2 \mid c_L)$	$\cdots$	$p(y_d \mid c_L)$

where each row corresponds to a possible conditioning context and each column corresponds to a possible value of the predicted variable. Since the probabilities in each row sum to one, each row contains $d - 1$ free parameters (since probabilities must sum to 1). Therefore, representing the full conditional distribution requires

$$L(d - 1)$$

free parameters.

Applying chain rule. First consider the case where the model has no information about the spectrum of the recording (i.e. no anchor). The unanchored joint is,

$$p(x) \;=\; \prod_{i=1}^{n} p\big(x_i \mid x_{<i}\big). \tag{5}$$

It must therefore be able to represent any distribution on $\mathcal{D}^n$, so it allocates a row in the counting table to every prefix $x_{<i}$. That is, $L = d^{i-1}$ and the number of parameters is

$$D_{\text{un}} \;=\; (d-1)\sum_{i=1}^{n} d^{\,i-1} \;=\; d^{\,n} - 1. \tag{6}$$

A.4 Anchored Parameter Count

Recall the set of band-limited signals $\mathcal{X}_K$ from equation 4. Let $\mathcal{K}$ be a family of frequency support patterns, each of size at most k. Let $\mathcal{X}_{\mathcal{K}} = \bigcup_{K \in \mathcal{K}} \mathcal{X}_K$, denoting the set of signals satisfying any allowed $\mathcal{K}$. We will argue that a band-limited signal may have n time-domain samples, but if it contains only $|K|$ active frequency components, it has only $|K|$ independent degrees of freedom.

Lemma A.1 (Band-limited signal identifiability)**.** *For every $K \subseteq \{0, \ldots, n-1\}$, the first $|K|$ samples of any $x \in \mathcal{X}_K$ determine x.*

Proof. Consider any $x \in \mathcal{X}_K$. By definition of $\mathcal{X}_K$, the Fourier coefficients outside K are zero:

$$\hat{x}_j = 0 \qquad \text{for all } j \notin K.$$

Thus, although the DFT contains n coefficients in total, only the $|K|$ coefficients $(\hat{x}_j)_{j \in K}$ are unknown. Since $\hat{x}_j = 0$ for $j \notin K$, the inverse DFT reduces to

$$x_t = \frac{1}{n} \sum_{j \in K} \hat{x}_j e^{2\pi i j t/n}, \qquad t = 0, \ldots, n-1. \tag{7}$$

Now restrict this equation to the first $|K|$ time samples, $t = 0, \ldots, |K| - 1$. This gives $|K|$ linear equations in the $|K|$ unknown Fourier coefficients $(\hat{x}_j)_{j \in K}$. Writing $K = \{j_1, \ldots, j_{|K|}\}$, these equations can be expressed as

$$\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{|K|-1} \end{bmatrix} = \frac{1}{n} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ e^{2\pi i j_1/n} & e^{2\pi i j_2/n} & \cdots & e^{2\pi i j_{|K|}/n} \\ (e^{2\pi i j_1/n})^2 & (e^{2\pi i j_2/n})^2 & \cdots & (e^{2\pi i j_{|K|}/n})^2 \\ \vdots & \vdots & \ddots & \vdots \\ (e^{2\pi i j_1/n})^{|K|-1} & (e^{2\pi i j_2/n})^{|K|-1} & \cdots & (e^{2\pi i j_{|K|}/n})^{|K|-1} \end{bmatrix} \begin{bmatrix} \hat{x}_{j_1} \\ \hat{x}_{j_2} \\ \vdots \\ \hat{x}_{j_{|K|}} \end{bmatrix}.$$

The middle matrix is a Vandermonde matrix (i.e. geometric progression in each column), so it is invertible. Hence the first $|K|$ samples uniquely determine all active Fourier coefficients $(\hat{x}_j)_{j\in K}$. Substituting the coefficients back into the inverse DFT in equation 7 determines x_t for every $t = 0, \dots, n-1$. Therefore, the first $|K|$ time samples uniquely determine the signal. □

Corollary 1 (Counting band-limited grid signals)**.**

$$|\mathcal{X}_K| \le d^{|K|} \qquad \textit{and} \qquad |\mathcal{X}_\mathcal{K}| \le |\mathcal{K}| d^k$$

Proof. We first count the number of distinct length-n time-domain signals that are possible under a particular bandwidth constraint. Since each time-domain sample takes a value in $\mathcal{D}$ and $|\mathcal{D}| = d$, there are at most $d^{|K|}$ possible tuples $(x_0, \dots, x_{|K|-1})$. Because each such tuple determines at most one signal in $\mathcal{X}_K$ by Lemma A.1, it follows that

$$|\mathcal{X}_K| \le d^{|K|}.$$

Now count the number of distinct signals that satisfy any bandwidth constraint in $\mathcal{K}$. Recall,

$$\mathcal{X}_\mathcal{K} = \bigcup_{K\in\mathcal{K}} \mathcal{X}_K.$$

Since every $K \in \mathcal{K}$ satisfies $|K| \le k$,

$$|\mathcal{X}_K| \le d^{|K|} \le d^k.$$

Therefore we can applied a union bound,

$$|\mathcal{X}_\mathcal{K}| = \left| \bigcup_{K\in\mathcal{K}} \mathcal{X}_K \right| \le \sum_{K\in\mathcal{K}} |\mathcal{X}_K| \le \sum_{K\in\mathcal{K}} d^k = |\mathcal{K}| d^k.$$

□

We formally define an anchor. While this work uses the spectrogram operator, we can consider the broader class of anchors given by deterministic functions of a signal that preserve its spectral support.

Assumption 1 (Band-limited spectral anchor)**.** *The anchor is $s = A(x)$ for a map A with $\operatorname{supp} A(x) = \operatorname{supp} \hat{x}$. The anchor alphabet contains only band-limited spectra, i.e., $\operatorname{supp} s \subseteq K$ for some $K \in \mathcal{K}$ whenever $s \in \mathcal{S}$. The joint is supported on consistent pairs (i.e. a signal cannot be paired with two different anchors): $p(x, s) > 0 \Rightarrow A(x) = s$.*

Assumption 1 requires the anchor to be computed from the signal itself, so it adds no randomness of its own. Thus, once x is known, so is s. Because $s_j = 0$ exactly when $\hat{x}_j = 0$, the anchor always reveals which frequency components are active, even when it leaves out details such as phase, and the band-limit constraint can be read directly from s. Every anchor corresponds to a signal satisfying one of the allowed bandwidth constraints in $\mathcal{K}$, and each signal is uniquely paired with its corresponding anchor. Conditioning on s therefore identifies which of the n frequency bins are active, while the anchor may contain varying amounts of information about their values. Both anchors we study below satisfy the assumption. The full spectrum $s_j = \hat{x}_j$ reveals the active bins together with their exact values, and the magnitude spectrum $s_j = |\hat{x}_j|$ reveals the active bins together with their energy only.

Under Assumption 1, the sequences compatible with an anchor value s form the set

$$\mathcal{X}(s) \;=\; \{\, x \in \mathcal{D}^n : A(x) = s \,\} \;\subseteq\; \mathcal{X}_{\operatorname{supp} s}. \tag{8}$$

The bandwidth constraint is therefore imposed through s, restricting the joint contexts $(s, x_{<i})$. We note that Assumption 1 holds for anchoring either with the full spectrum, $s_j = \hat{x}_j$ or the magnitude spectrum, $s_j = |\hat{x}_j|$, (and possibly quantized).

Lemma A.2 (Complexity of band-limited anchoring)**.** *Under Assumption 1: The sets $\mathcal{X}(s)$, $s \in \mathcal{S}$, are disjoint subsets of $\mathcal{X}_{\mathcal{K}}$. Consequently,*

$$\sum_{s \in \mathcal{S}} |\mathcal{X}(s)| \leq |\mathcal{K}| d^k.$$

Proof. In this proof, we are upper bounding the number of distinct possible anchors. Because the anchor is a deterministic function of the signal, each signal x has exactly one anchor $A(x)$. Therefore, for any two distinct anchors $s \neq s'$, $\mathcal{X}(s) \cap \mathcal{X}(s') = \emptyset$. Moreover, by Assumption 1, every anchor $s \in \mathcal{S}$ has spectral support contained in some $K \in \mathcal{K}$. By Corollary 1,

$$\sum_{s \in \mathcal{S}} |\mathcal{X}(s)| = \left| \bigcup_{s \in \mathcal{S}} \mathcal{X}(s) \right| \leq |\mathcal{X}_{\mathcal{K}}| \leq |\mathcal{K}| d^k.$$

□

We previously showed that representing the unanchored distribution $p(x)$ requires $D_{\text{un}} = d^n - 1$ free parameters. We now introduce the deterministic anchor $s = A(x)$ and count the free parameters required to represent the joint distribution $p(x, s) = p(s)p(x \mid s)$. To keep the notation simple, we assume that every recording compatible with an anchor occurs with positive probability, i.e., $p(x, s) > 0$ for every $s \in \mathcal{S}$ and $x \in \mathcal{X}(s)$. (Otherwise, replace $\mathcal{X}(s)$ by the recordings that actually occur with s; all upper bounds below are unchanged.)

Proposition 1 (Support-aware anchored count)**.** *Under Assumption 1,*

$$D_{\text{anch}}^{\text{supp}} = \underbrace{\big(|\mathcal{S}| - 1\big)}_{\text{anchor } p(s)} + \underbrace{\sum_{s \in \mathcal{S}} \big(|\mathcal{X}(s)| - 1\big)}_{\text{generator } p(x|s)} = \left| \bigcup_{s \in \mathcal{S}} \mathcal{X}(s) \right| - 1 \leq |\mathcal{K}| d^k - 1. \tag{9}$$

Proof. For each $s \in \mathcal{S}$, the conditional distribution $p(x \mid s)$ is supported on $\mathcal{X}(s)$ and therefore requires $|\mathcal{X}(s)| - 1$ free parameters. The anchor distribution $p(s)$ requires $|\mathcal{S}| - 1$ free parameters. Hence,

$$D_{\text{anch}}^{\text{supp}} = \big(|\mathcal{S}| - 1\big) + \sum_{s \in \mathcal{S}} \big(|\mathcal{X}(s)| - 1\big) = \sum_{s \in \mathcal{S}} |\mathcal{X}(s)| - 1.$$

Because the anchor map A is deterministic, the sets $\mathcal{X}(s)$ are disjoint. Therefore,

$$\sum_{s \in \mathcal{S}} |\mathcal{X}(s)| = \left| \bigcup_{s \in \mathcal{S}} \mathcal{X}(s) \right|.$$

Finally, Lemma A.2 gives

$$\left| \bigcup_{s \in \mathcal{S}} \mathcal{X}(s) \right| \leq |\mathcal{K}| d^k,$$

which proves the result. □

In the unanchored case, we naively allocate parameters to all d^n possible time-domain sequences as in the unstructured representation of $p(x)$. In Equation 9, we show that the anchored factorization $p(x, s) = p(s)p(x \mid s)$ restricts the parameterization to the set of feasible signals, which is substantially smaller due to the bandwidth constraint. Consequently, the anchored factorization yields a lower parameter complexity than an unstructured time-domain parameterization. Note that the parameter complexity bound is determined by the bandwidth constraints rather than the choice of anchor. The choice of anchor instead determines how this complexity is allocated between the anchor model and the conditional generator, as described in the following remark.

Remark 1 (Effect of anchor informativeness)**.** *The choice of anchor determines how the support-aware parameter complexity is divided between the anchor model $p(s)$ and the conditional generator $p(x \mid s)$.*

- *At one extreme, a full complex-spectrum anchor determines x through the inverse DFT, so $|\mathcal{X}(s)| = 1$ for every s and the conditional generator contributes no free parameters. All parameter complexity is therefore allocated to $p(s)$.*
- *At the other extreme, a support-only anchor reveals only which frequency bins are active, leaving the corresponding spectral values unresolved and more variability to $p(x \mid s)$.*
- *A magnitude-based spectral anchor lies between these extremes, resolving spectral magnitudes while leaving information such as phase to the conditional generator.*

A.5 Specializing to a Contiguous Bandwidth Model

We now express these bounds in terms of physical quantities to illustrate the change in parameter complexity between unanchored and anchored representations. Suppose a signal of bandwidth B is sampled at rate f_s for n samples, i.e., for a duration $T = n/f_s$. Each DFT bin then spans f_s/n Hz, so the signal occupies about $(B/f_s)\, n$ of the n frequency bins. Since the signal may sit anywhere in the band, we allow every contiguous block of this width.

Corollary 2 (Gain under a contiguous bandwidth constraint)**.** *We specialize the allowed-support family $\mathcal{K}$ to all circular windows of k contiguous DFT bins. We assume $d \geq 2$ and $n \geq 2$.*

For any anchor satisfying Assumption 1, the support-aware parameter count in equation 9 satisfies

$$\log \frac{D_{\text{un}}}{D_{\text{anch}}^{\text{supp}}} \geq \left(1 - \frac{B}{f_s}\right) n \log d - \log(2nd) = (f_s - B)T \log d - \log(2nd). \tag{10}$$

Proof. We specialize Lemma A.2 and Proposition 1 to the contiguous bandwidth family defined above. Because the occupied band may begin at any of the n DFT bins, $|\mathcal{K}| = n$. Moreover, $k = \left\lceil \frac{B}{f_s} n \right\rceil \leq \frac{B}{f_s} n + 1$.

The unstructured time-domain parameterization is

$$D_{\text{un}} = d^n - 1 \geq \frac{d^n}{2}.$$

By Proposition 1, restricting the parameterization to signals consistent with the allowed bandwidth supports gives

$$D_{\text{anch}}^{\text{supp}} \leq |\mathcal{K}| d^k.$$

Substituting $|\mathcal{K}| = n$ and the bound on k yields

$$D_{\text{anch}}^{\text{supp}} \leq n d^{(B/f_s)n+1}.$$

Therefore,

$$\frac{D_{\text{un}}}{D_{\text{anch}}^{\text{supp}}} \geq \frac{d^n/2}{n d^{(B/f_s)n+1}} = \frac{1}{2nd} d^{(1-B/f_s)n}.$$

Taking logarithms gives

$$\log \frac{D_{\text{un}}}{D_{\text{anch}}^{\text{supp}}} \geq \left(1 - \frac{B}{f_s}\right) n \log d - \log(2nd).$$

Finally, using $n = f_s T$ gives

$$\left(1 - \frac{B}{f_s}\right) n = (f_s - B)T,$$

which yields the second form. □

Remark 2 (Bandwidth Gain Regimes)**.** *The gain exponent in equation 10 is the bandwidth that the anchor rules out, $f_s - B$, multiplied by the duration T. Equivalently, each of the $n - k \approx (f_s - B)\, T$ frequency bins that the anchor forces to zero divides the parameter count by a factor of d. Three regimes follow:*

- Narrowband *($B \ll f_s$): the gain is nearly d^n. The extreme case is a DC signal, where the only allowed support is $K = \{0\}$, so $|\mathcal{K}| = 1$ and $k = 1$. Then $\mathcal{X}_{\{0\}}$ consists of the d constant sequences $x = c\mathbf{1}$ with $c \in \mathcal{D}$, and, given the anchor, x_0 fixes every later sample. Proposition 1 gives $D_{\text{anch}}^{\text{supp}} \leq d - 1$, which does not grow with n at all, compared with $D_{\text{un}} = d^n - 1$.*
- Fixed fraction *($B/f_s = \beta < 1$): the gain grows exponentially in the duration, as $d^{(1-\beta)n}$ up to polynomial factors.*
- Full band *($B = f_s$): here $k = n$, and the bound guarantees no gain. This is expected because, without a band limit, every sequence in $\mathcal{D}^n$ is possible, and the anchor rules nothing out.*

B Extended Related Works

Multimodal IQ Embeddings. IQFM uses self-supervised pretraining on IQ streams and LoRA adaptation for downstream tasks including modulation classification and RF fingerprinting (Mashaal & Abou-Zeid, 2026). Other approaches exploit complementary wireless representations: IQFormer fuses IQ and time-frequency features for modulation recognition, while multimodal wireless foundation models learn across IQ, spectrograms, channel state information, and channel impulse responses (Shao et al., 2025; Aboulfotouh & Abou-Zeid, 2025). However, these methods are optimized for downstream discrimination and do not natively support conditioning variables for controlled signal generation across modalities. In particular, we use CLIP-style contrastive alignment (Radford et al., 2021) between signal embeddings and semantic metadata alongside reconstruction, producing a decodable, label-aware latent space for conditional waveform generation as shown in Section 5.1.

Conditional Time Series and Audio Generation Time Weaver incorporates heterogeneous metadata, including categorical, continuous, and time-varying variables, into a conditional diffusion model (Narasimhan et al., 2024). WaveStitch additionally uses partially observed signals at inference and reconciles overlapping, parallel-generated windows through a stitching objective (Shankar et al., 2025). RF-Diffusion modifies the signal diffusion process itself to operate across time and frequency (Chi et al., 2024). In adjacent waveform domains, DiffWave conditions audio synthesis on a mel spectrogram, whereas AudioLDM generates in a compressed mel-spectrogram latent space and decodes the result to audio (Kong et al., 2021; Liu et al., 2023a). These approaches use frequency information as a generation condition or within the generative process. ALF instead uses a separately learned spectrogram latent to supervise an intermediate flow representation during training; the target spectrogram is not supplied at inference.

Intermediate Representation Supervision. Anchored Diffusion Language Models predict distributions over important tokens and use those predictions to guide subsequent denoising (Rout et al., 2025a). In image generation, REPA instead aligns intermediate denoiser representations with features from a pretrained encoder of the clean image (Yu et al., 2025). ALF uses a complementary spectral representation to expose structure that may be difficult to learn directly in the IQ generation space. Specifically, a training-time spectral anchor encourages intermediate flow representations to capture global time-frequency structure associated with the target waveform. The spectral target is then removed during generation, avoiding extra computational cost at inference.

Generative Channel Reconstruction. Another branch of work models the wireless channel $\mathbf{H}$, which characterizes propagation in the simplified received-signal relation $\mathbf{Y} = \mathbf{H} * \mathbf{X} + \mathbf{N}$. Here, $\mathbf{X}$ is the transmitted waveform and $\mathbf{N}$ is receiver noise; in channel estimation, the transmitted pilot portion of $\mathbf{X}$ is known. Generative estimators reconstruct $\mathbf{H}$ from such measurements using priors learned by Wasserstein GANs, score-based models, or diffusion models (Balevi et al., 2020; Arvinte & Tamir, 2022; Zhou et al., 2025). For example, the GAN prior was evaluated on $\mathbf{H} \in \mathbb{C}^{16\times 64}$, while a subsequent score-based method evaluated matrices up to $\mathbb{C}^{64\times 256}$ (Balevi et al., 2020; Arvinte & Tamir, 2023). For channel synthesis, location-conditioned DDIM generates $\mathbb{C}^{4\times 32}$ beamspace channels from three-dimensional user positions, while a physics-informed VAE generates channels through a parametric geometric model (Lee et al., 2026; Wagle et al., 2025). These approaches support applications such as beam alignment, channel compression, and data

augmentation. Their channel-focused outputs do not, however, represent the additional variability of the transmitted waveform and receiver noise in complete received recordings $\mathbf{Y} \in \mathbb{C}^{4\times N}$ covered in ALF.

C Dataset Design and Validation

Here we discuss the detailed implementation and properties of the WaveScene dataset.

C.1 Physical Channel Model and Modeling Challenges

For a clean complex-baseband waveform $x[n]$, the received signal at antenna a is

$$y_a[n] = \sum_{\ell=1}^{L} h_{a,\ell} x[n - d_\ell] + w_a[n], \qquad a \in \{1, \ldots, 4\}, \tag{11}$$

where $h_{a,\ell}$ is the antenna-dependent complex coefficient of path ℓ, d_ℓ is its shared discrete delay, and $w_a[n]$ is independent complex receiver noise. The four receive elements observe the same transmitted waveform, path delays, and propagation geometry, while their complex path coefficients vary according to the array response. Their spatial dependence can be represented by

$$\mathbf{R}_{h,\ell} = \mathbb{E}\left[\mathbf{h}_\ell \mathbf{h}_\ell^H\right], \qquad \mathbf{h}_\ell = [h_{1,\ell}, h_{2,\ell}, h_{3,\ell}, h_{4,\ell}]^T. \tag{12}$$

Consequently, a recording cannot be reduced to four independent SISO observations: relative phase, amplitude, covariance, and shared multipath structure carry information about the propagation environment.

All recordings use a single transmit antenna and a 4-element horizontal uniform linear receive array with half-wavelength spacing. The element spacing is rescaled for each carrier frequency. The array elements are vertically polarized and omnidirectional. UMi, UMa, and RMa channels are generated with Sionna v1.2.1 (Hoydis et al., 2023), using its implementation of the geometry, large-scale parameters, visibility conditions, path loss, and shadow fading specified by 3GPP TR 38.901 (3GPP, 2024b). InH channels are generated with QuaDRiGa v2.8.1 (Jaeckel et al., 2014), using the `3GPP_38.901_Indoor_Open_Office` scenario and a corresponding four-element horizontal receive array.

The corpus presents two complementary modeling challenges. First, recordings contain as many as 32,768 complex samples, requiring preservation of waveform timing, spectral structure, and protocol- or radar-specific behavior over long contexts. Note that we use a static channel impulse response (CIR), so temporal evolution of a mobile channel is out of scope. Second, the four synchronized receive streams must remain spatially consistent: a model must preserve cross-antenna phase, amplitude, and covariance relationships in addition to producing realistic marginal signals.

C.2 Waveform Parameterization and Standards Alignment

The pre-channel waveform library used in WaveScene is constructed around documented physical-layer timing, bandwidth, and carrier constraints instead of arbitrary signal perturbations that may conflict with technical documentations shown in Table 5. It contains 1,066 parameterized waveform variants from 32 classes spanning generic modulations, communications protocols, and radar emissions. Toolbox- backed WLAN, Bluetooth, LTE, and NR classes use standards-oriented physical-layer implementations. Other classes preserve their documented timing, bandwidth, pulse, hopping, and signal-morphology assumptions, although we do not uniformly claim receiver-decodable bitstreams for these classes.

Table 5: Representative physical assumptions used in waveform generation.

Family	Representative choices	Physical basis
Linear modulation	RRC shaping with roll-off $\alpha \in \{0.20, 0.25, 0.35\}$; varied symbol and capture rates	Digital modulation; DVB-S2
WLAN	20-MHz PPDU configurations, MCS values, preambles, and guard intervals	IEEE 802.11 PHYs
Bluetooth	LE 1M/2M GFSK and BR/EDR 2-Mb/s $\pi/4$-DQPSK PHYs	Bluetooth Core radio PHY
LTE, NR, and NB-IoT	OFDM numerologies, resource grids, and physical-channel configurations	3GPP LTE and NR PHYs
Zigbee	2-Mchip/s O-QPSK with 32-chip DSSS symbol morphology	IEEE 802.15.4 PHY
LoRa	CSS with SF6–SF9 and 125/250/500-kHz bandwidths	Semtech LoRa PHY
ATSC	8-VSB segment timing, pilot, synchronization, and vestigial-sideband morphology	ATSC A/53 Part 2
Radar	Pulse, PRI, chirp, hop, dwell, and phase-code parameterizations	Classical radar waveform texts

Sources. Linear modulation: (Proakis & Salehi, 2008; ETSI, 2014); WLAN: (IEEE, 2020a; 2021); Bluetooth: (Bluetooth SIG, 2023); LTE, NR, and NB-IoT: (3GPP, 2023; 2024a); Zigbee: (IEEE, 2020b); LoRa: (Semtech Corporation, 2020); ATSC: (Advanced Television Systems Committee, 2011); radar: (Skolnik, 2008; Richards et al., 2010; Levanon & Mozeson, 2004).

For waveform classes with a configured bandwidth B_{cfg}, the release checker enforces

$$B_{99} \le \rho B_{\mathrm{cfg}}, \qquad B_{\mathrm{cfg}} \le f_s, \tag{13}$$

where B_{99} is the measured 99%-power occupied bandwidth, ρ is a class-specific tolerance, and f_s is the complex-baseband sampling rate. The occupied band is obtained from a Welch power spectral density estimate (Welch, 1967) using the interval between the 0.5% and 99.5% cumulative-power quantiles. This definition accommodates asymmetric spectra such as SSB-AM and ATSC 8-VSB.

Sampling rate, bandwidth, and recording length are selected jointly. WLAN is generated at its native 20-MHz rate and, where applicable, at valid oversampled rates; LoRa uses capture rates appropriate for its 125/250/500-kHz bandwidths; narrowband analog signals use correspondingly lower rates; and radar sampling rates scale with chirp bandwidth or hop span. For pulsed and frequency-hopping radar, occupied bandwidth is measured over the active pulse or hop sequence rather than over silent receive intervals. This prevents waveform duty cycle from being mistaken for narrow spectral occupancy.

Controlled payload content. Payload bits are deliberately controlled rather than independently randomized across recordings. A deterministic canonical bit stream is generated once, and the modulation and protocol waveforms draw length-appropriate prefixes from this stream before class-specific framing, coding, scrambling, and modulation. This choice prevents arbitrary payload content from becoming an additional unlabeled source of variation, allowing the corpus to emphasize waveform morphology and propagation effects. Consequently, the dataset does not assess generalization to independently varying payload content; extending the corpus with independently seeded payloads is a natural direction for future work.

C.3 Propagation-State Construction of Shared-Channel Set

For the 7.52M shared-channel set, Propagation states are intentionally not balanced independently of scene because their frequencies form part of the assumed deployment model (3GPP, 2024b). For urban microcell (UMi) and urban macrocell (UMa), 80% of generated links are initialized as outdoor-to-indoor (O2I) and 20% as outdoor. O2I forces an indoor receiver, whereas outdoor links are assigned line-of-sight (LOS) or non-line-of-sight (NLOS) status by Sionna's scene and distance-dependent visibility model. We use the LOS/NLOS proportions from the realized geometries instead of forcing a balanced distribution. Indoor hotspot (InH) links contain no O2I state; their LOS or NLOS condition is generated by the QuaDRiGa Indoor Open Office model.

Rural macrocell (RMa) requires an explicit approximation. The documented deployment mixture (3GPP, 2024b) contains 50% indoor and 50% in-car users, but Sionna does not provide a vehicle-

cabin penetration model. Due to this limitation, we approximate by retaining the indoor half as O2I and representing the in-car half as outdoor without vehicle penetration loss. Sionna then divides this outdoor component into LOS and NLOS links according to the realized geometry. These leads to the empirical distributions shown in Table 6 and Figure 7:

Table 6: Empirical propagation-state composition of the 7.52M shared-channel set.

Scene	LOS (%)	NLOS (%)	O2I (%)
UMi	8.0	11.9	80.0
UMa	5.1	14.9	80.0
RMa	19.2	30.0	50.8
InH	68.1	31.9	–

C.4 Label Definitions and Normalization

The 6 learning factors are derived from the clean waveform manifests and generated channel realizations rather than estimated from the final noisy IQ. Waveform class, scene type, and realized propagation state are categorical. The continuous targets are spectral occupancy, mean per-antenna pre-combining SNR, and RMS delay spread.

Spectral occupancy is defined as

$$o = \frac{B_{99}}{f_s}, \tag{14}$$

where B_{99} is the measured 99%-power occupied bandwidth, and f_s is the native complex-baseband sampling rate. Unlike an absolute bandwidth label, o describes the fraction of the sampled spectrum occupied by the waveform and is therefore comparable across capture rates.

The reported SNR is the mean per-antenna pre-combining SNR:

$$\mathrm{SNR}_{\mathrm{dB}} = P_{\mathrm{tx,dBm}} + G_{\mathrm{path,dB}} - P_{\mathrm{noise,dBm}}. \tag{15}$$

Here, $P_{\mathrm{tx,dBm}}$ is transmit power, $G_{\mathrm{path,dB}}$ is channel path gain obtained by averaging CIR energy across the four receive antennas, and $P_{\mathrm{noise,dBm}}$ is the receiver noise power. Noise power is calculated as

$$P_{\mathrm{noise,dBm}} = -174 + 10\log_{10}(B_{99}) + 10\log_{10}(T/290) + \mathrm{NF}, \tag{16}$$

with temperature $T = 290$ K and noise figure $\mathrm{NF} = 9$ dB. The four streams contain independent AWGN, and no coherent combining gain is included in the learning target.

RMS delay spread is calculated from the power-weighted discrete CIR:

$$\tau_{\mathrm{rms}} = \sqrt{\sum_{\ell=1}^{L} w_\ell(\tau_\ell - \bar{\tau})^2}, \qquad \bar{\tau} = \sum_{\ell=1}^{L} w_\ell \tau_\ell, \qquad w_\ell = \frac{|h_\ell|^2}{\sum_{j=1}^{L}|h_j|^2}. \tag{17}$$

In this expression, τ_ℓ is the delay of tap ℓ, h_ℓ is its complex coefficient after averaging tap energy across receive antennas, and w_ℓ is its normalized power contribution.

The continuous targets are normalized as

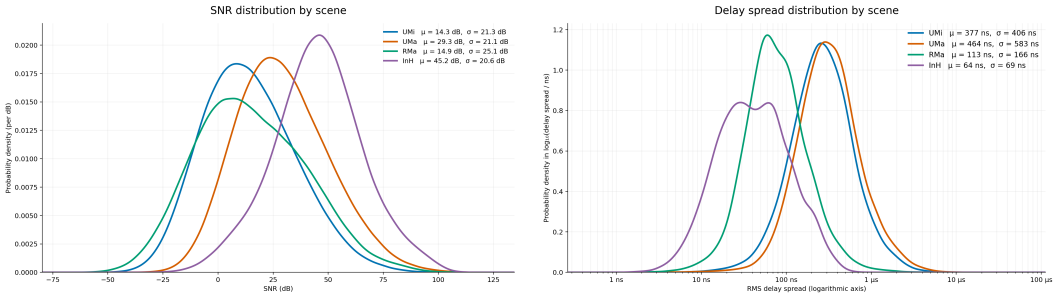
$$\widetilde{\mathrm{SNR}} = \mathrm{clip}\left(\frac{\mathrm{SNR}_{\mathrm{dB}} - 25}{25}, -3, 3\right), \tag{18}$$

$$\tilde{o} = \mathrm{clip}\left(\frac{\log_{10}(o) + 1}{0.75}, -3, 3\right), \tag{19}$$

$$\tilde{\tau} = \mathrm{clip}\left(\frac{\log_{10}(\tau_{\mathrm{rms}}/\mathrm{ns}) - 2.4}{0.8}, -3, 3\right). \tag{20}$$

Clipping is rare in the 7.52M-recording corpus. For SNR, 0.018% of recordings fall below the lower threshold and 0.102% exceed the upper threshold. No occupancy values are clipped. For delay spread, 0.0014% fall below the lower threshold and fewer than 0.0001% exceed the upper threshold.

Carrier frequency, transmit power, horizontal distance, path gain, channel identity, and diagnostic channel quantities are retained as metadata despite not being used as primary learning factors. These fields remain available for auditing and stratified analysis when needed.



(a) SNR (b) RMS delay spread

Figure 7: Empirical SNR and RMS delay-spread distributions by propagation scene of the shared-channel set.

C.5 Held-Out Channel Benchmark

The held-out-channel benchmark contains 300,000 recordings whose channel identities do not occur in the development corpus. Channel identity jointly includes scene, topology, stochastic channel variation, propagation state, and carrier frequency. Holding out this complete identity prevents the benchmark from evaluating a previously observed CIR under only a different waveform or noise realization.

The benchmark contains exactly 9,375 recordings from each of the 32 waveform classes. It is also well distributed across propagation scenes and carrier frequencies, while retaining the physically motivated state composition described above. Its purpose is to measure generalization to unseen propagation realizations, but it does not provide a deliberately shifted waveform or label distribution. The deterministic list of selected recording and channel identifiers is distributed with the benchmark.

C.6 Reproducibility and Limitations

Dataset construction is controlled by versioned waveform manifests and generation plans. Each waveform manifest specifies its physical parameters, sampling rate, stored length, active region, padding, and measured occupied bandwidth. A master seed is deterministically expanded into topology, state-selection, waveform-selection, channel, and noise seeds using SHA-256-derived identifiers. Each channel cache stores a fingerprint tied to its waveform-slot plan, generation configuration, geometry rules, channel-engine version, array configuration, and cache schema. The pipeline rejects caches whose fingerprints do not match the active configuration.

The release checker validates all 1,066 waveform manifests and generated pre-channel waveforms. It verifies identifier agreement, stored lengths, normalization, padding constraints, occupied-bandwidth conditions, and SHA-256 waveform uniqueness. We will release the dataset-generation code, configuration manifests, seed policies, label definitions, and validation scripts upon paper acceptance.

Although WaveScene covers diverse waveform families and propagation conditions, it has several limitations. Channels remain static within each recording, so the current release does not include mobility, Doppler evolution, or handover. We align waveforms to a normalized temporal origin; asynchronous detection and crop localization therefore fall outside the benchmark. The receiver model includes thermal noise and receiver noise figure, but excludes carrier-frequency offset, sampling-clock offset, IQ imbalance, nonlinear hardware distortion, co-channel interference, and external atmospheric or man-made noise. For RMa, the nominal in-car condition uses a Sionna-compatible approximation: half of these links are modeled as O2I and half as outdoor links, with Sionna sampling LOS or NLOS for the outdoor portion. All selected carriers lie within the 0.5–100 GHz scope of TR 38.901. Communications payloads use a deterministic reference source, as described above, so the benchmark does not evaluate generalization to independently varying message content. Finally, several manually implemented protocol classes preserve documented RF timing and waveform morphology without providing complete receiver-decodable protocol stacks.

The radar classes model emitted or intercepted waveforms and exclude two-way target propagation, range-dependent return delay, radar cross section, target Doppler, and clutter.

D Encoder Architecture and Evaluation Details

D.1 Encoder Architecture

This section provides implementation details ommited from Section 5.1. Stage 1 of ALF uses $N_{\text{patch}} = 1024$ and a 768-dimensional latent for each patch. The raw-IQ and ButterflyNet-inspired spectral branches are fused by a learned gate; the spectral branch is initialized with a radix-2 Fourier inductive bias. A 4-layer, 8-head Transformer mixes information across patches to form per-patch embedding $\mathbf{Z}_{\text{IQ}} \in \mathbb{R}^{P\times 768}$, and a `[CLS]` token produces per-recording embedding $\mathbf{Z}_{\text{IQ-CLS}} \in \mathbb{R}^{768}$. A separate 4-layer, 8-head decoder Transformer first mixes the complete latent sequence across patches. Each resulting token is then decoded by a dual-path patch decoder: a transposed-convolutional raw-IQ path provides the primary waveform reconstruction, while an inverse-Butterfly path provides a gated spectral correction. A separate linear head predicts the patch log-RMS amplitude, so received level does not dominate waveform-shape reconstruction.

Reconstruction objective. Let $\widehat{\mathbf{X}}_{\text{IQ}}$ denote the decoded recording. The reconstruction loss is

$$\begin{aligned}\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{IQ}} + \mathcal{L}_{\text{corr}} + 0.5\mathcal{L}_{\text{FFT}} + \mathcal{L}_{\text{STFT}} \\ + 0.5\mathcal{L}_{\text{env}} + 0.25\mathcal{L}_{\text{cov}} + 0.25\mathcal{L}_{\text{level}}.\end{aligned} \tag{21}$$

Here, $\mathcal{L}_{\text{IQ}}$ is a sample-wise ℓ_1 loss on the real–imaginary waveform components, and $\mathcal{L}_{\text{corr}}$ is one minus the cosine similarity of the flattened reconstructed and target IQ waveforms. The latter is scale-invariant but phase-sensitive, preventing a low-amplitude output from always achieving a low loss.

The remaining terms compare complementary signal properties: $\mathcal{L}_{\text{FFT}}$ is an ℓ_1 loss between full-patch $\log(1+|\text{FFT}(\cdot)|^2)$ spectra; $\mathcal{L}_{\text{STFT}}$ compares log-power spectrograms using a 256-point STFT with hop size 64; and $\mathcal{L}_{\text{env}}$ compares magnitude envelopes pooled into 32 temporal windows. To preserve SIMO spatial structure, $\mathcal{L}_{\text{cov}}$ is the mean absolute discrepancy between trace-normalized complex antenna covariance matrices. Finally, $\mathcal{L}_{\text{level}}$ is an ℓ_1 loss on the scalar amplitude predicted by the decoder's level path.

Soft-target contrastive alignment. For a batch of B recordings, let $\mathbf{z}_i$ and $\mathbf{c}_j$ be the normalized IQ and label embeddings defined in Section 5.1, and let

$$\ell_{ij} = \frac{\mathbf{z}_i^\top \mathbf{c}_j}{\tau}, \qquad \tau = 0.2. \tag{22}$$

Rather than using a one-hot pairing target, we construct a metadata affinity matrix. Let w_i, s_i, and r_i denote waveform class, scene type, and propagation state; let a_i, ρ_i, and d_i denote spectral occupancy, mean per-antenna pre-combining SNR in dB, and RMS delay spread in ns, respectively. We use

$$\begin{aligned}m_{ij} = \mathbb{1}[w_i = w_j]\,\mathbb{1}[s_i = s_j]\,\mathbb{1}[r_i = r_j] \\ \times \exp\left[-\frac{1}{2}\left(\left(\frac{\rho_i - \rho_j}{3}\right)^2 + \left(\frac{\log_{10} a_i - \log_{10} a_j}{0.15}\right)^2 + \left(\frac{\log_{10} d_i - \log_{10} d_j}{0.25}\right)^2\right)\right].\end{aligned} \tag{23}$$

The normalized row targets are $q_{ij} = m_{ij}/\sum_{k=1}^{B} m_{ik}$. The symmetric SoftCLIP loss is

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2B}\sum_{i=1}^{B}\sum_{j=1}^{B}\left[q_{ij}\log \text{softmax}_j(\ell_{ij}) + q_{ji}\log \text{softmax}_j(\ell_{ji})\right]. \tag{24}$$

Consequently, recordings must agree on the three discrete label fields to be positive pairs, while nearby occupancy, SNR, and delay-spread values receive larger relative weight. This avoids treating semantically similar recordings as equally negative merely because they have varied labels.

Structural objective. The last loss applied to $\mathbf{Z}_{\text{IQ-CLS}}$ is the structure loss, $\mathcal{L}_{\text{struct}}$.It encourages the representation to retain each metadata factor, distribute factors across distinct embedding groups, and preserve nonzero variation across embedding dimensions.

Specifically, the 768-dimensional embedding is partitioned into $G = 8$ groups of 96 dimensions. For each of the six metadata fields

$$\mathcal{F} = \{\text{waveform}, \text{occupancy}, \text{SNR}, \text{scene}, \text{propagation state}, \text{delay spread}\},$$

a lightweight gated MLP probe predicts the corresponding target. Categorical factors use cross-entropy and continuous factors use mean-squared error after the training-set normalization. Denoting the sum of these six losses by $\mathcal{L}_{\text{probe}}$, the complete structural loss is

$$\begin{aligned}\mathcal{L}_{\text{struct}} = \alpha_{\text{probe}}\mathcal{L}_{\text{probe}} + \alpha_{\text{gate}}\mathcal{L}_{\text{gate}} + \alpha_{\text{group}}\mathcal{L}_{\text{group}} \\ + \alpha_{\text{dim}}\mathcal{L}_{\text{dim}} + \alpha_{\text{var}}\mathcal{L}_{\text{var}}.\end{aligned} \tag{25}$$

Each factor has a softmax-normalized gate over the eight groups. $\mathcal{L}_{\text{gate}}$ penalizes overlap between the average gates for the waveform–occupancy factors and those for the propagation-related factors (SNR, scene, propagation state, and delay spread). This encourages, but does not force, the two types of information to use distinct portions of the embedding.

For a minibatch embedding matrix $\mathbf{Z} \in \mathbb{R}^{B\times 768}$, $\mathcal{L}_{\text{group}}$ is the mean squared cross-covariance between distinct 96-dimensional groups, while

$$\mathcal{L}_{\text{dim}} = \frac{1}{768}\sum_{u\neq v}\text{Cov}(\mathbf{Z})_{uv}^2, \qquad \mathcal{L}_{\text{var}} = \frac{1}{768}\sum_{u=1}^{768}\left[\max(0, \gamma - \text{Std}(\mathbf{Z}_{:u}))\right]^2, \tag{26}$$

with $\gamma = 0.02$. The group and dimension covariance terms discourage redundant encoding, whereas the variance-floor term directly guards against dimensional collapse. The gated factor probe is used only as a training-time auxiliary and is discarded after Stage 1; it is distinct from the post-hoc MLP probes used for evaluation.

Optimization schedule. We jointly optimize $\mathcal{L}_{\text{S1}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{CLIP}}(e)\mathcal{L}_{\text{CLIP}} + \lambda_{\text{struct}}(e)\mathcal{L}_{\text{struct}}$ for 20 epochs using AdamW with learning rate 2×10^{-4}, weight decay 0.05, and gradient clipping at 1.0. Reconstruction and the initial structural terms are active from the outset. The CLIP coefficient is scheduled from 0 to 0.25 and then 0.5, while the structural coefficients are increased from their initially weak settings over later epochs. This staging allows the decoder to first establish a stable reconstructive latent before strong alignment and decorrelation pressures are applied.

Training and probing protocol. Stage 1 of ALF is trained on the deterministic training partition of the WaveScene 5.6M development set, comprising approximately 5.3M recording–label pairs for training; the remaining 5% of that development set (0.3M pairs) is used for validation. The seen and unseen-channel test sets later discussed are never used during training. We train for 20 epochs with AdamW, a learning rate of 2×10^{-4}, weight decay of 0.05, gradient clipping at 1.0, and a global batch size of 288 across 4 GPUs. All Stage-1 components, including the IQ encoder, decoder, and label encoder, are subsequently frozen.

The representation probes reported in this work are post-hoc evaluations and are distinct from the training-time factor probes in $\mathcal{L}_{\text{struct}}$. For each frozen encoder and prediction task, we train an identical one-hidden-layer MLP on a fixed 5M subset of the 5.3M training partition, select the model using a disjoint 256K validation set, and repeat training with 5 fixed random seeds. Each resulting probe is evaluated unchanged on both a 300K-recording subset of the complete 1.9M seen-channel test set and the 300K channel-held-out test set. The seen-channel subset is matched to the channel-held-out set across waveform class, scene type, and propagation state, isolating the effect of channel realization while preserving the WaveScene label space.

D.2 Adaptation of External RF Encoders

We compare ALF with WavesFM (Aboulfotouh et al., 2025), IQFM (Mashaal & Abou-Zeid, 2026), and IQFormer (Shao et al., 2025). Our WaveScene recordings are longer and span a broader combination of waveform families, channel environments, and propagation states than the data used by

the external encoders in their original publications. We therefore regard these as best reproducible adaptations of the available baseline implementations: they preserve each method's native encoder and validated input recipe while matching the WaveScene training recordings and frozen-probe protocol wherever retraining is possible. We do not add ALF's cross-patch mixing, label encoder, reconstruction objective, or structural constraints to any external baseline.

WavesFM. We use the released pretrained checkpoint, retaining its IQ tokenizer and 8-block Transformer. The encoder receives four-antenna, $N_{\text{view}} = 256$ IQ views with per-recording maximum-absolute normalization, and its 256-dimensional `[CLS]` feature is used for probing. For deterministic evaluation, we use the center view from the fixed window bank. No WaveScene waveform, scene, or channel label is used to train or adapt this encoder.

IQFM. We retain IQFM's native multi-antenna SSL-Joint ShuffleNetV2-x0.5 encoder and its self-supervised objective. Each recording contributes one $N_{\text{view}} = 256$ view per training visit, selected from up to 8 evenly spaced views; the two contrastive augmentations are constructed from that selected view. Evaluation uses the deterministic center view. Thus, the training budget counts source recordings rather than treating multiple views from one recording as independent recordings.

IQFormer. We retain IQFormer's published single-antenna IQ–STFT encoder and remove its classification layer to expose the 64-dimensional pooled feature. Its time–frequency preprocessing uses a Blackman-window STFT with window length 31, hop 1, FFT size 128, and the first 32 frequency bins. IQFormer is trained with waveform-class supervision over the 32 WaveScene waveform classes before probe training; therefore, its waveform result should be interpreted with the favorable supervised pretraining setting explicitly in mind. We do not use its unpublished auxiliary power feature.

ALF encoder. Also refered to as ALF's stage 1, ALF encoder processes the complete native recording as a sequence of 1,024-sample patches and uses the frozen $\mathbf{Z}_{\text{IQ-CLS}}$ embedding from Stage 1 training (5.1). Its encoder therefore has access to recording-level context unavailable to external baselines due to their different $N_{\text{view}} = 256$.

Training comparability. Stage 1 ALF training uses the established 5.36M-recording training partition. The retrained IQFM and IQFormer encoders use a fixed 5M-recording training subset with a disjoint 256K-recording validation subset. All downstream frozen-probe and label-efficiency evaluations use the same fixed 5M-recording fit subset, disjoint 256K-recording validation subset, and matched test sets for every encoder. Runtime efficiency is measured at the same input lengths for every model. Baseline-specific augmentation, supervision, and optimization details are retained when they are intrinsic to the published method. WavesFM instead uses its released pretrained checkpoint. The external baselines' native 256-sample views are a limitation of this comparison; we preserve them to avoid silently replacing their validated training regimes with a new long-context architecture.

Encoder size and long-recording scaling. Tables 7 and 8 compare the encoder-only model size and batch-1 inference cost. For ALF, we exclude all reconstruction-only and generation-only components, including the decoder-side cross-patch mixer, dual-path waveform decoder, amplitude head, label encoder, spectral autoencoder, and latent flow. ALF therefore uses a 60.31M-parameter encoder to produce its 768-dimensional recording-level embedding.

We measure FP32 CUDA-event latency and peak forward activation memory on an NVIDIA H100 NVL using inputs of $N_{\text{samples}} \in \{2048, 8192, 32768\}$; data loading and downstream MLP probes are excluded. ALF increases from 9.6 ms to 12.6 ms across this length range, whereas IQFormer increases from 6.4 ms to 89.8 ms. The external encoders are evaluated through their implemented variable-length forwards for this scaling diagnostic, although their reported frozen-probe results use their native 256-sample views. Consequently, Table 8 characterizes long-recording computational scaling rather than an equal-context representation comparison.

Table 7: Encoder-only model size.

Encoder	Parameters (M)	FP16 weights (MiB)	Embedding dim.
ALF	**60.31**	**115.0**	**768**
WavesFM	6.40	12.2	256
IQFM	0.34	0.7	1024
IQFormer	0.35	0.7	64

Table 8: Encoder-only batch-1 FP32 inference cost on one GPU. Each model uses its native antenna convention; ALF, WavesFM, and IQFM receive 4 antennas, while IQFormer receives antenna 0. Values exclude resident model weights, data loading and downstream heads.

Encoder	$N_{\text{samples}} = 2{,}048$		$N_{\text{samples}} = 8{,}192$		$N_{\text{samples}} = 32{,}768$	
	Latency (ms) ↓	Peak activation (MiB) ↓	Latency (ms) ↓	Peak activation (MiB) ↓	Latency (ms) ↓	Peak activation (MiB) ↓
ALF	**9.6**	**10.6**	**11.2**	**42.5**	**12.6**	**170.0**
WavesFM	1.9	3.4	1.9	3.5	2.0	4.3
IQFM	2.8	1.1	2.8	4.4	2.7	9.8
IQFormer	6.4	41.1	23.9	68.3	89.8	173.4

D.3 Frozen-Encoder Probing Protocol

All encoders are frozen before downstream probing. For every encoder and target, we train the same one-hidden-layer MLP,

$$\mathbf{h}_i = \mathrm{GELU}(\mathbf{W}_1 \overline{\mathbf{z}}_i + \mathbf{b}_1), \qquad \widehat{\mathbf{y}}_i = \mathbf{W}_2 \,\mathrm{Dropout}_{0.1}(\mathbf{h}_i) + \mathbf{b}_2, \tag{27}$$

where $\overline{\mathbf{z}}_i$ is the feature standardized using statistics computed only from the probe-training partition, and the hidden width is 512.

Each probe is trained on 5M recordings from the fixed training subset and selected using the disjoint 256K-recording validation set. We use AdamW with learning rate 10^{-3}, weight decay 10^{-2}, batch size 16,384, and 5 epochs. For every encoder–task pair, we train 5 matched seeds $(13, 29, 42, 71, 101)$ and retain the checkpoint with the best validation macro-F1 for classification or lowest validation RMSE for regression. The same saved probe is evaluated without adaptation on both the matched 300K-recording seen-channel test subset and the 300K channel-held-out test set.

Waveform recognition covers all 32 classes and is summarized by macro-F1. Mean per-antenna pre-combining SNR is evaluated as continuous regression in dB, while RMS delay spread is regressed in $\log_{10}(\mathrm{ns})$. Accuracy, macro recall, AUROC, MAE, and rank correlation are retained as secondary metrics. Table 9 reports mean performance and two-sided 95% Student-t confidence intervals across the 5 matched seeds.

Table 9: Frozen-encoder probing results. Waveform scores are macro-F1 (%); SNR is regressed in dB; and delay spread is regressed in $\log_{10}(\mathrm{ns})$. Values are mean $\pm$ two-sided 95% Student-t confidence intervals over 5 matched probe seeds.

Encoder	Seen channels			Channel-held-out channels		
	Waveform F1 (%) ↑	SNR RMSE ↓	Delay RMSE ↓	Waveform F1 (%) ↑	SNR RMSE ↓	Delay RMSE ↓
WavesFM	50.35 ± 0.08	15.937 ± 0.024	0.4822 ± 0.0007	51.08 ± 0.08	15.348 ± 0.016	0.4784 ± 0.0009
IQFormer	66.18 ± 0.09	9.416 ± 0.023	0.4482 ± 0.0003	67.11 ± 0.13	9.062 ± 0.014	0.4419 ± 0.0005
IQFM	60.00 ± 0.55	14.173 ± 0.051	0.4294 ± 0.0013	60.64 ± 0.51	13.608 ± 0.036	0.4243 ± 0.0012
ALF	$\mathbf{99.21 \pm 0.01}$	$\mathbf{1.259 \pm 0.004}$	$\mathbf{0.1752 \pm 0.0001}$	$\mathbf{99.37 \pm 0.01}$	$\mathbf{1.201 \pm 0.004}$	$\mathbf{0.1678 \pm 0.0001}$

D.4 Label efficiency

To measure how readily task information can be extracted from each frozen representation, we repeat waveform probing with nested deterministic prefixes of 1K, 10K, 100K, 1M, and 5M recordings from the same ordered probe-training split. The validation set, probe architecture, optimizer, training duration, five seeds, and 300K test sets remain unchanged. Even the 1K prefix contains all 32 waveform classes, with 22–40 recordings per class.

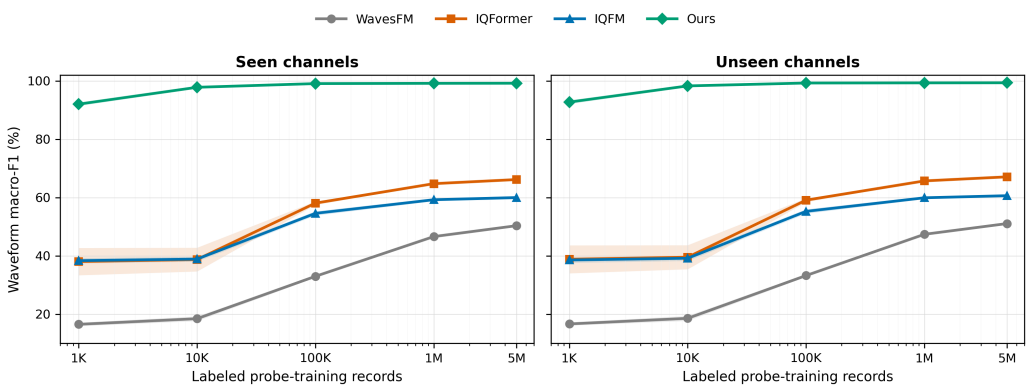


Figure 8: Waveform classification as the number of labeled probe-training recordings is varied. Curves show mean macro-F1 and shaded 95% confidence intervals across five matched seeds. Our frozen representation reaches 92.0% and 92.7% macro-F1 with only 1K labeled recordings on seen and unseen channels, respectively, and exceeds 99% by 100K recordings.

The nearly coincident seen and channel-held-out curves show that ALF's label efficiency extends beyond the channel realizations used for training. Because WaveScene training data already span substantial channel variability, the channel-held-out set evaluates controlled channel generalization, not a severe distribution shift. Our encoder reaches 97.8% seen-channel and 98.3% unseen-channel macro-F1 with 10K recordings and largely saturates by 100K. In comparison, the external encoders continue improving between 100K and 5M recordings but plateau at substantially lower performance. IQFormer's wider confidence interval at 1K and 10K additionally indicates greater sensitivity to the small probe-training set.

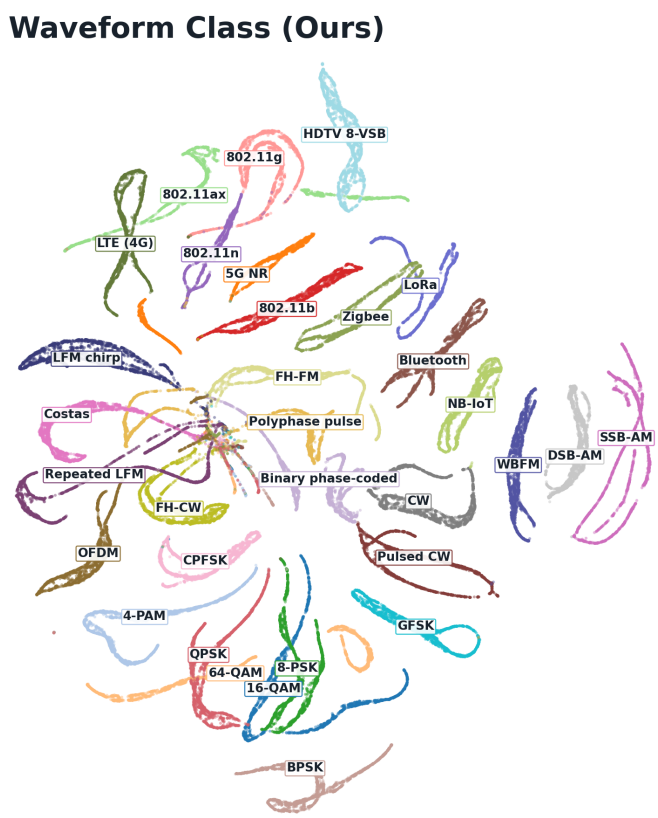
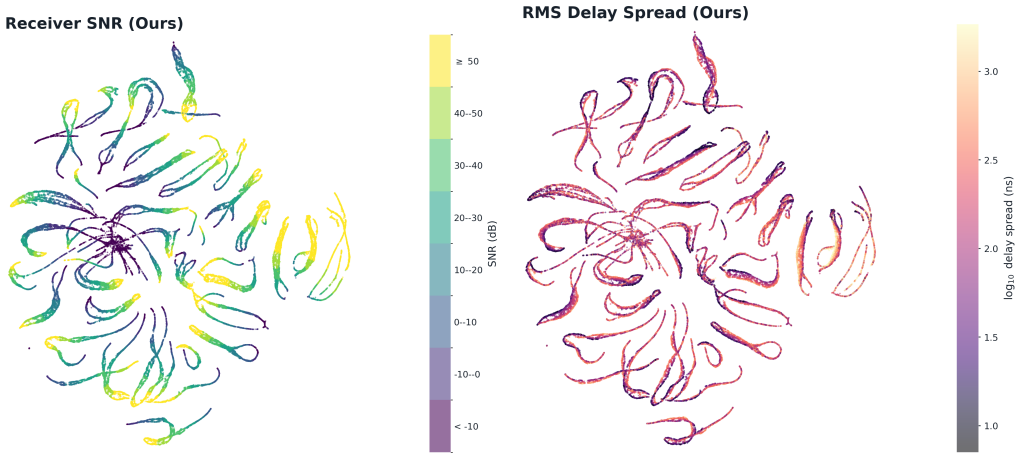


Figure 9: UMAP of the ALF encoder's signal embedding over waveform class.

D.5 Embedding geometry



(a) Receiver SNR.

(b) RMS delay spread.

Figure 10: ALF encoder's UMAP coordinates colored by receiver SNR and RMS delay spread.

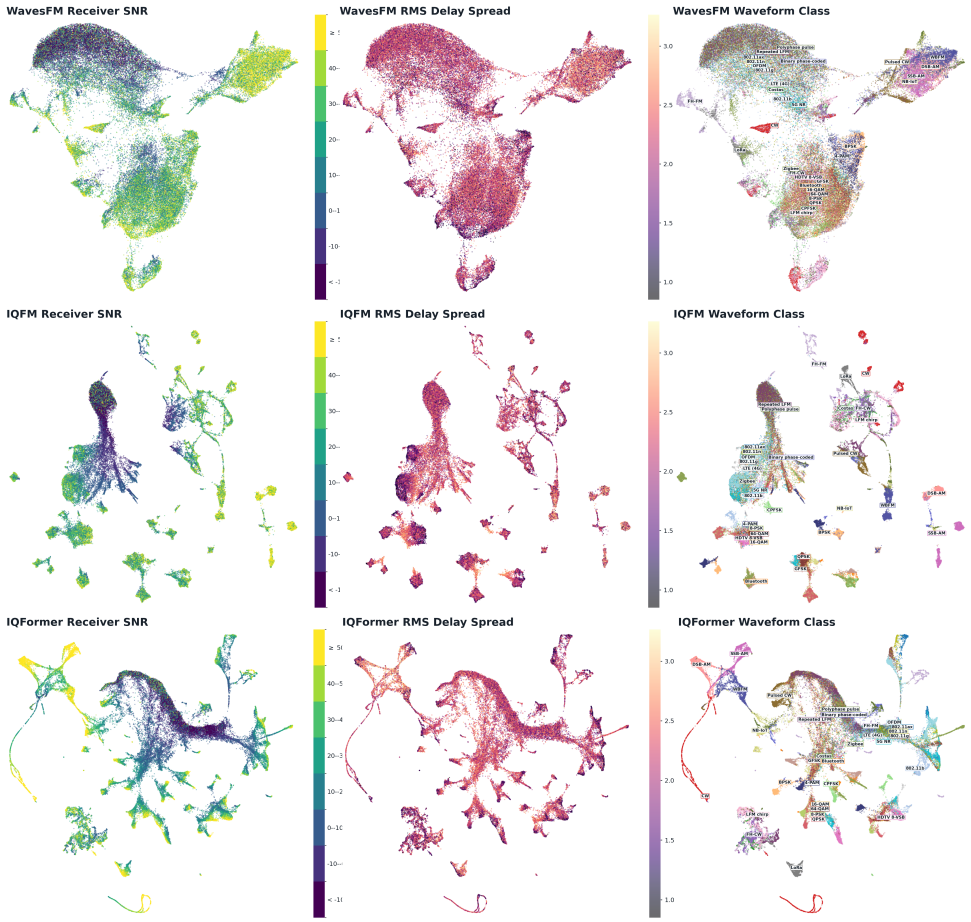


Figure 11: Frozen-encoder UMAPs for WavesFM, IQFM, and IQFormer. Within each row, waveform, receiver-SNR, and delay-spread panels share the same coordinates.

We visualize the final 768-D signal embedding on 64,000 deterministically selected held-out seen-channel recordings using UMAP (McInnes et al., 2018). Embeddings are first normalized to unit length and reduced to 50 principal components, which retain 45.1% of their variance. UMAP is then fitted with 50 neighbors, minimum distance 0.15, cosine distance, and random seed 42. All panels use the same recordings and coordinates, only the displayed labels change.

Figures 9 and 10 show that ALF's final embedding organizes waveform class into distinct manifolds while retaining smooth variation in mean per-antenna pre-combining SNR and RMS delay spread. In contrast, the adapted external encoders exhibit more overlapping qualitative geometry in Figure 11. As with any UMAP visualization, these plots are descriptive; the frozen-encoder probe results (Table 9) provide the quantitative comparison.

D.6 Zero-shot alignment and exact retrieval

Beyond reconstruction, we evaluate whether the frozen encoder is directly useful for downstream RF understanding. These experiments test its label-aligned embedding space without adapting the signal encoder: conditional zero-shot prediction measures attribute-level organization, while exact retrieval measures whether a signal can recover its complete metadata configuration.

For conditional zero-shot prediction, we construct a metadata prototype for each candidate target value. For waveform classification, the candidates are the 32 waveform classes. For SNR and occupancy, we partition the training-set values into 32 quantile bins and use one prototype per bin. A prototype is the average metadata embedding from 512 training recordings having the corresponding waveform class or continuous-value bin; their remaining metadata attributes retain their natural training-set variation. At test time, we assign each signal to the prototype with the largest cosine

similarity. For continuous targets, the prediction is the training-set median of the selected bin. Test labels are used only to score these predictions.

Exact paired retrieval tests a stricter form of alignment. For each signal query, we form its positive candidate from the complete 6-label metadata configuration of the same recording: waveform class, scene type, propagation state, mean per-antenna pre-combining SNR, spectral occupancy, and RMS delay spread. We rank this candidate against the metadata configurations of every recording in the same test gallery using cosine similarity. Recall at K (R@K) is the fraction of queries for which the matched 6-factor configuration appears among the K highest-ranked candidates.

Table 10: Zero-shot alignment and exact paired retrieval for the proposed encoder. Conditional zero-shot values are mean $\pm$ 95% confidence interval over five prototype-sampling seeds.

Metric	Seen channels	Unseen channels
Zero-shot waveform accuracy (%) $\uparrow$	95.55 ± 0.21	95.87 ± 0.23
Zero-shot SNR RMSE (dB) $\downarrow$	5.204 ± 0.213	4.811 ± 0.176
Zero-shot occupancy RMSE (log fraction) $\downarrow$	0.151 ± 0.009	0.134 ± 0.010
Exact retrieval R@1 (%) $\uparrow$	28.17	27.72
Exact retrieval R@5 (%) $\uparrow$	60.02	59.62
Exact retrieval R@10 (%) $\uparrow$	72.84	72.66

Each retrieval gallery contains 300K candidate metadata configurations, so every query is ranked against a large set of closely related negative recordings that may share waveform, channel attributes, or both. Despite this strict matched-gallery setting, the encoder retrieves the complete 6-factor configuration within its top 10 candidates for 72.84% of seen-channel queries and 72.66% of channel-held-out queries, demonstrating that its aligned embedding preserves substantially more than waveform identity alone.

E ALF DETAILS AND EVALUATIONS

E.1 HYPERPARAMETERS

Table 11: ALF hyperparameters. Values marked † are inherited from the frozen autoencoder and are not free parameters of the flow.

Hyperparameter	Value	Role
Backbone		
Depth	12	Number of DiT blocks (Peebles & Xie, 2023)
Width	768	Residual and token width of every DiT block
Heads	12	Attention heads, head dimension 64
MLP expansion	4	Hidden width of each block's feed-forward layer
Max patch tokens	32	Longest record the model emits in one pass
Patch length	$1024^{\dagger}$	IQ samples per patch
Patch dimension	$768^{\dagger}$	Latent width per patch
Flow parameters	131,134,464	Size of the flow network, excluding the anchor head, the frozen autoencoder, and the conditioning encoder.
Conditioning		
Conditioning dimension	768	Width of the label vector
Conditioning factors	6	Waveform class, scene type, propagation state, estimated SNR, occupancy, RMS delay spread § C.4
Label encoder	frozen	The encoder is fixed
Prior	enabled	Sampling from label encoder mean and variance
Latent preprocessing and objective		
Objective	$v = Z_1 - Z_0$	Rectified-flow velocity target
Standardizer batches	60	Batches per rank used to fit the per-dimension latent mean and variance
Clip σ	8.0	Winsorising threshold applied after standardising
Time distribution	logit-normal $(0, 1)$	Distribution the flow time t is drawn from during training
Signal scale	10^4	Fixed input scaling the encoder was fitted under.
Optimization		
Optimizer	AdamW	-
Learning rate	1×10^{-4}	Peak rate.
Weight decay	0.02	Applied to trainable flow parameters only
Gradient clip	1.0	Global gradient-norm clip applied before each step.
Batch size	64×4	Records per rank across four ranks; effective batch size 256.
Epochs	20	Full passes over the training split.
Schedule	warmup + cosine	Stepped once per epoch: one warmup epoch, then cosine decay. The rate is flat at peak for the first two epochs and at $\sim 0.7\%$ of peak at the end.
Seed	42	Initialization and data order.
Anchor		
Anchor layer	6 of 12	At the midpoint of DiT Depth
Anchor hidden width	768	Width of the head's MLP
Anchor weight	0.25	Weight of the auxiliary term relative to the flow loss
Hold / ramp epochs	0 / 2	Weight ramps linearly
Time weighting	$t/\bar{t}$	Weights each record's anchor loss toward low t
Head parameters	2,363,904	Discarded at inference, only used for anchoring
Sampling		
ODE steps	50	Integration steps from noise to data.
Integrator	Heun	Second-order predictor–corrector solver.

Hyperparameter	Value	Role
Guidance scale	1.0	No classifier-free guidance; conditioning enters only through the modulation path.
Output normalisation	RMS	Rescales each record to unit mean power
Records per corpus	100,000	Generated under held-out test-split conditioning.
Sampling seed	0	Sampling noise draw.
Edge refiner	≥ 30 dB	Refined variants only: a post-hoc pass over records whose conditioning SNR clears this gate. Unrefined corpora are otherwise identical.

E.2 Label-Conditioned Gaussian Prior

Rather than transporting isotropic white Gaussian noise $\mathbf{Z}_0 \sim \mathcal{N}(0, \mathbf{I})$ to clean latents, ALF initializes its rectified-flow trajectories from the multimodal source distribution $p_{\text{src}}(\cdot \mid \mathbf{Y})$ parameterized by the frozen probabilistic label tower $\theta_{\text{lbl}}(\mathbf{Y})$. The label encoder consumes exactly the six physical RF metadata factors $\mathbf{Y}$—waveform class, SNR, channel occupancy, propagation scenario, channel state, and RMS delay spread (Appendix C.4)—mapping them into a 768-dimensional diagonal Gaussian. For an arbitrary record of length P patches (spanning 2,048 to 32,768 IQ samples at 1,024 samples per patch), the initial flow state at $t = 0$, denoted $\mathbf{Z}_0 \in \mathbb{R}^{P \times 768}$, is constructed by taking P independent draws from that record's conditioning cloud:

$$\mathbf{Z}_0^{(p)} = \boldsymbol{\mu}_{\text{lbl}}(\mathbf{Y}) + \boldsymbol{\sigma}_{\text{lbl}}(\mathbf{Y}) \odot \mathbf{z}_p, \quad \mathbf{z}_p \sim \mathcal{N}(0, \mathbf{I}), \quad p \in \{1, \ldots, P\}. \tag{28}$$

Because the rectified flow interpolation $\mathbf{Z}_t = (1-t)\mathbf{Z}_0 + t\mathbf{Z}_1$ and velocity target $v = \mathbf{Z}_1 - \mathbf{Z}_0$ require only that training and sampling draw from the identical distribution, this formulation restricts the flow's starting domain to a class-structured hyperspherical shell ($\|\boldsymbol{\mu}_{\text{lbl}}(\mathbf{Y})\| \approx 24.9$ versus $\sqrt{768} \approx 27.7$) retaining approximately 43% of standard normal entropy per dimension ($\sigma_{\text{lbl}} \approx 0.444$). Under this label-conditioned source, conditioning dropout is strictly set to 0.0; an unconditional branch would otherwise default to standard normal noise, invalidating the velocity-mixing assumption of classifier-free guidance. Conditioning information is thus present twice: statically as the spatial boundary state $\mathbf{Z}_0$ of the flow trajectory, and dynamically across all flow times t via adaptive LayerNorm (adaLN) modulation in the DiT backbone, which jointly injects the normalized direction of $\boldsymbol{\mu}_{\text{lbl}}(\mathbf{Y})$ alongside a learned patch-length embedding of P.

E.3 Separate Spectral Autoencoder and Mid-Depth Anchoring

To enforce physically consistent spectral properties without incurring test-time computational overhead, ALF incorporates the auxiliary mid-depth supervisory anchor driven by the independently trained, frozen spectral autoencoder (38.6M parameters). Operating on per-patch STFT log-power spectrograms (256 frequency bins $\times$ 13 time frames per patch, computed with an FFT size and window of 256, hop size 64, and antenna power averaging), this autoencoder is trained with absolute level preservation, ensuring the latent representation faithfully encodes both spectral shape and wide dynamic range (spanning -15 dB to $+68$ dB SNR). Its post-mixing latent sequence $\mathbf{Z}_{\text{spec}}$ reflects global recording context via a 4-layer Transformer mixer while maintaining local invertibility back to the patch spectrum via a per-patch 2D convolutional decoder.

During flow training, the spectral autoencoder remains frozen. The auxiliary prediction head θ_{anchor} ($768 \to 768 \to 768$ MLP, 2.36M parameters) is tapped off the intermediate hidden representation $\mathbf{H}_t^{(6)}$ after the 6th Transformer block of the 12-block flow backbone. The head acts as a mid-depth denoising probe trained to predict the standardized clean spectrogram latent $\widetilde{\mathbf{Z}}_{\text{spec}}$ via mean squared error. Velocity matching is deliberately avoided because coupling the IQ latent noise draw with an independently trained spectral latent space is ill-posed. As shown in Equation 2, the auxiliary spectral loss is weighted by $t/\bar{t}$:

$$\mathcal{L}_{\text{anchor}} = \frac{t}{\bar{t}} \left\| \theta_{\text{anchor}}\left(\mathbf{H}_t^{(6)}\right) - \widetilde{\mathbf{Z}}_{\text{spec}} \right\|_2^2, \tag{29}$$

which concentrates gradient pressure toward the target IQ distribution ($t \to 1$) where intermediate activations carry recoverable spectral structure. The anchor term enters the total objective with

weight $\lambda_{\text{anchor}}(e)$ ramped linearly to 0.25 over the first two epochs, with its output projection zero-initialized so as not to destabilize early flow optimization. Because θ_{anchor} provides representational regularization exclusively during backpropagation, it is discarded at inference, adding zero parameters and zero FLOPs to generation.

E.4 Training Consistency and Generation Protocol

To isolate the empirical gains of the inductive biases during ablation, all flow models were trained to a convergence under identical optimization protocols: AdamW with decoupled weight decay (0.02), gradient-norm clipping at 1.0, a batch size of 256 across 4 ranks, and a 20-epoch warmup-cosine learning rate schedule peaking at 10^{-4} (with checkpoints selected by lowest held-out validation loss; epoch 19 for ALF). For downstream evaluation, all synthetic corpora were generated under an identical protocol: 100,000 records conditioned on the held-out test split under fixed sampling seed 0, integrated from $t = 0$ to $t = 1$ using a second-order Heun ODE solver at 50 steps, guidance scale 1.0 (no classifier-free guidance), decoded to time-domain waveforms using the frozen decoder $\theta_{\text{dec-IQ}}$, and scaled via per-record RMS power normalization to unit mean energy (see Table 11 for full hyperparameter specifications). The 5.6M development set with the same deterministic train/validation is used for training.

The three baselines differ in how far they support the corpus's 5 recording lengths ($N_{\text{samples}} \in \{2048, 4096, 8192, 16384, 32768\}$). **TimeWeaver** (Narasimhan et al., 2024) is natively variable-length: a single model is trained jointly on all five buckets because nothing in the denoiser is learned per length. Batches are length-homogeneous (bucket sampling, with no padding or attention mask avoids feeding masked zeros into the target), and the per-rank batch is scaled as $16 \cdot 2048/L$ so that $B \cdot C \cdot L$ is held constant; together with the sum-reduced training loss, this makes the loss scale length-invariant, so the original learning rate transfers mostly unchanged. **WaveStitch** (Shankar et al., 2025) covers the same range only by tiling: its S4 backbone (Gu et al., 2022) sees a fixed 256-sample window, so a length-L record is cut into $L/256$ non-overlapping windows, the same conditioning vector is broadcast to every window, each window is denoised independently, and the results are concatenated. We generate in the metadata-only mode, which uses neither stitching nor self-guidance, so the windows are uncoupled: the model has no mechanism for structure longer than one window, and a seam occurs every 256 samples. **RF-Diffusion** (Chi et al., 2024) is not adapted at all—it applies full $O(N^2)$ self-attention to raw samples with no patching, which is tractable at 2048 but not at 8192 and above without an architectural change, so it is both trained and evaluated at $L = 2048$ only. Its numbers therefore correspond to the 2048 column and are not comparable to the other arms' length-pooled figures. Again, the 5.6M development set with the same deterministic train/validation is used for training, and we trained till a convergence is reached.

E.5 Inference

The cost of a multi-stage anchored flow training finally yields its benefits at inference. Pure data generation is divided into three pieces: Sample, Flow, Decode (Figure 12). The only user-specified information needed is the desired features that characterize the target waveform. Additionally, our per-patch design (§ 5.1,5.2) makes inference length-agnostic, with potentially no limit on generation length (Appendix F). Label specification conditions the flow model's prior, giving $p_{z(1)}(z|c)$. ALF continues to pay dividends; its latent nature achieves significant speed advantages to prior works (Table 4, § 6).

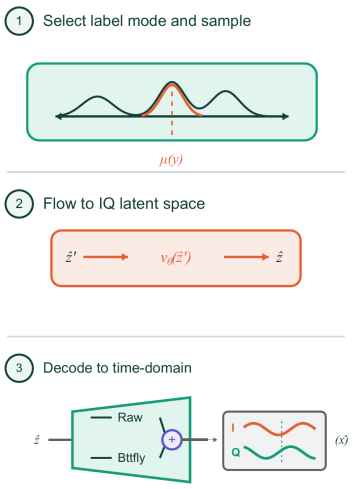


Figure 12: Inference Pipeline

E.6 Refined vs. Unrefined

Due to the nature of the per-patch AE structure, although trained for continuity, some results still show small fluctuations at patch seams. In the frequency domain, these high-frequency components translate to stripes in generations. Although signal integrity is preserved (i.e., overall SNR is the same), these artifacts resulted as increase in MSE metrics (PSD and Spectrogram). Thus, we built a refining mechanism for post-decode smoothing.

The refiner is a small dilated 1-D residual CNN with per-antenna inputs I, Q, $|z|$, and the cosine and sine of the lag-one phase increment; the mixer tokens of the two patches adjoining each join, (h_{p-1}, h_p), are linearly projected and added to the feature map within ± 32 samples of that join. This network emits a per-antenna polar correction $(\Delta a, \Delta\phi)$, $\tanh$-bounded to ± 0.5 in log-amplitude and $\pm\pi/2$ in phase, applied multiplicatively as $z \leftarrow z\, e^{m\Delta a} e^{jm\Delta\phi}$. We use a polar correction instead of a complex linear blend because blending phase-decorrelated signals can create amplitude notches and broadband artifacts. For 802.11ax ($\kappa \approx 0.15$), scaling and rotation avoid these issues. At inference, an SNR threshold is set to avoid tampering with intentionally noisy generations, and records above the determined level are passed through this refinement pipeline to smooth seam transitions.

Table 12: MSE within 95% Confidence over 100k Recordings

	Unrefined		Refined	
Run	Med PSD MSE	Med Spec MSE	Med PSD MSE	Med Spec MSE
Baseline w/o Prior	19.77 ± 0.20	127.85 ± 0.48	17.87 ± 0.16	120.29 ± 0.50
Baseline w/ Prior	19.17 ± 0.18	128.71 ± 0.50	17.48 ± 0.15	121.09 ± 0.47
ALF (ours)	$\mathbf{18.52 \pm 0.19}$	$\mathbf{126.54 \pm 0.46}$	$\mathbf{16.86 \pm 0.14}$	$\mathbf{119.08 \pm 0.45}$
ALF vs. Base w/o Prior	+6.3%	+1.0%	+5.6%	+1.0%
ALF vs. Base w/ Prior	+3.4%	+1.7%	+3.5%	+1.7%

Table 12 shows the distributional improvements of roughly 2dB and 6.5dB for PSD and Spectrogram MSEs, respectively, across all ablations. A design choice was made to include this refiner for all latent flow models (ALF, LF-M, LF-N) compared in the main paper (Table 3,§ 6).

E.7 Robustness

Table 3 shows that ALF slowed the record collapse curve. Figure 13 visualizes the reclassification data in Table 13 and improvements, including a broader view across inference steps. Notably, the same trend continues through 5 steps; at a single inference step, ALF remains the best-performing. Finally, note that Time Weaver experiences the largest collapse in this region due to the naturally noisier process of diffusion models over flows.

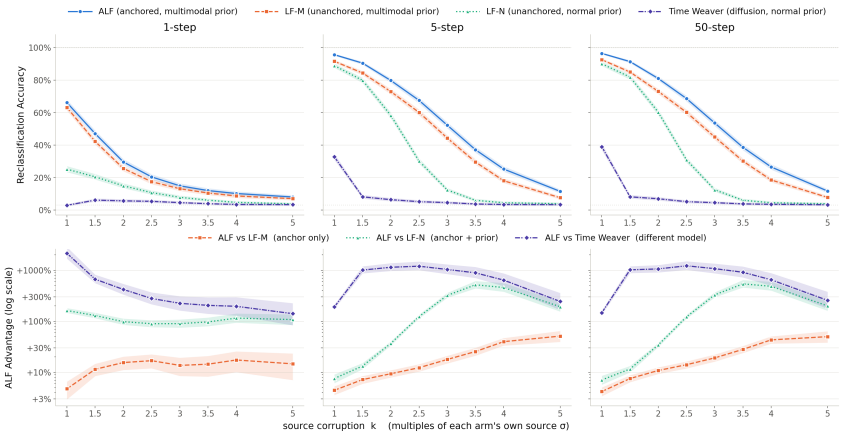


Figure 13: Signal reclassification accuracies for low-density sampling and various inference steps

Table 13: Recovery under source corruption at matched sigma multiples $\mathbf{k}\sigma$

	5 flow steps							50 flow steps						
	Multimodal Gaussian			Standard normal				Multimodal Gaussian			Standard normal			
k	**ALF**	LF	$\Delta^{\mathrm{LF}}_{\%}$	LF	TW	$\Delta^{\mathrm{LF}}_{\%}$	$\Delta^{\mathrm{TW}}_{\%}$	**ALF**	LF	$\Delta^{\mathrm{LF}}_{\%}$	LF	TW	$\Delta^{\mathrm{LF}}_{\%}$	$\Delta^{\mathrm{TW}}_{\%}$
1.0	**95.7**	91.6	**+4.5%**	88.9	32.8	**+7.6%**	**+192.1%**	**96.5**	92.6	**+4.2%**	90.1	39.0	**+7.1%**	**+147.3%**
1.5	**90.5**	84.4	**+7.2%**	79.8	8.2	**+13.3%**	**+1009.2%**	**91.4**	85.0	**+7.6%**	81.9	8.2	**+11.6%**	**+1020.7%**
2.0	**79.8**	72.9	**+9.4%**	58.3	6.4	**+36.9%**	**+1145.4%**	**81.0**	73.1	**+10.9%**	60.3	7.0	**+34.4%**	**+1062.3%**
2.5	**67.5**	60.1	**+12.3%**	30.2	5.2	**+123.8%**	**+1202.1%**	**68.7**	60.2	**+14.1%**	30.9	5.2	**+122.2%**	**+1223.8%**
3.0	**52.3**	44.3	**+17.9%**	12.4	4.6	**+321.8%**	**+1037.8%**	**53.6**	45.0	**+19.2%**	12.5	4.6	**+329.5%**	**+1075.3%**
3.5	**37.1**	29.6	**+25.6%**	6.0	3.8	**+519.0%**	**+890.4%**	**38.7**	30.2	**+28.2%**	6.0	3.8	**+540.9%**	**+913.9%**
4.0	**25.2**	18.0	**+39.9%**	4.5	3.4	**+459.2%**	**+641.3%**	**26.5**	18.5	**+43.3%**	4.5	3.5	**+485.9%**	**+651.8%**
5.0	**11.5**	7.6	**+51.2%**	3.9	3.3	**+192.9%**	**+244.9%**	**11.7**	7.8	**+50.0%**	3.9	3.3	**+200.0%**	**+257.1%**

Note. LF-M: unanchored LF with a metadata prior; LF-N: unanchored LF with a standard-normal prior. TimeWeaver (TW) uses equal-NFE DDIM sampling with 10 and 100 steps, respectively.

E.8 Style Transfer Cont.

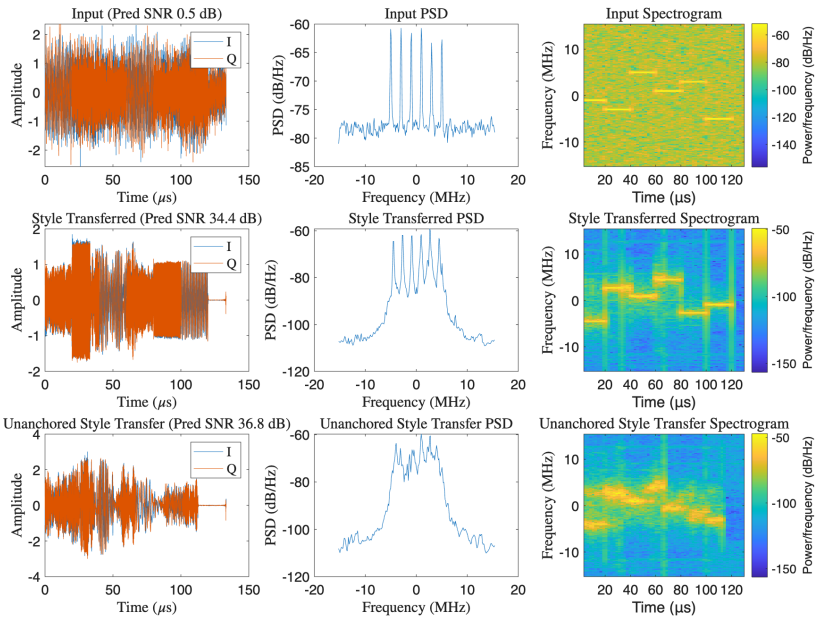


Figure 14: Unanchored style transfers are more prone to spectral blurring without anchored guidance

Signals are encoded, and their modes are decided through a max likelihood defined as:

$$\hat{c} \;=\; \arg\max_{c\in\mathcal{C}} \frac{1}{P}\sum_{p=1}^{P} \log \mathcal{N}\big(z_0^{(p)} \,\big|\, \mu_c,\, \mathrm{diag}(\sigma_c^2)\big) \tag{30}$$

We also highlight that anchoring reduces the observed blurring effect in the unanchored result. The following uses naive sampling while ablating ALF and LF-M (unanchored metadata-prior) models. As seen below, the unanchored style transfer's spectrogram has far less sharp frequency detail. This is expected because anchoring constrains flowed latent trajectories during style-transfer regeneration (Appendix A).

E.9 Within-Recording Phase Consistency

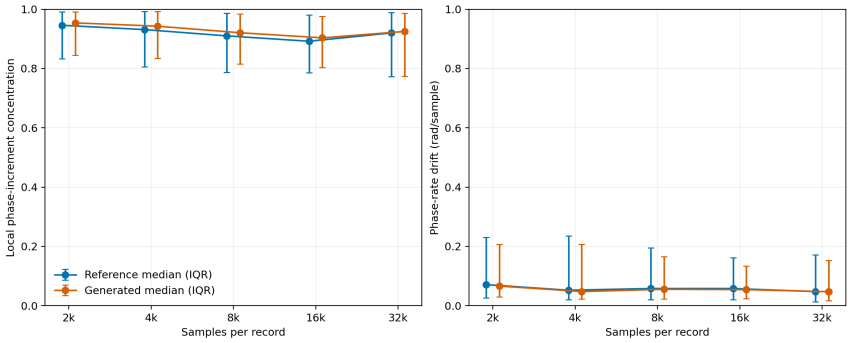


Figure 15: Within-record phase behavior by waveform length. Points show medians and bars show interquartile ranges across records. Left plot measures the consistency of adjacent-sample phase changes within normalized windows; higher is better. Right plot shows phase-rate drift, which measures variation of the window-level phase increment across each record; lower is better.

Phase encodes timing and carrier-related structure essential for coherent wireless reception; even small local phase errors can disrupt demodulation and cross-antenna coherence, making phase preservation a stringent test of long-recording generation. We evaluated whether ALF preserves within-recording phase behavior using 100,000 generated recordings conditioned on a subset of the channel-held-out test-set conditions and compared them with the corresponding reference recordings. We retained the 71,673 recordings with stored mean per-antenna pre-combining $\mathrm{SNR} > 10$ dB, as lower SNR substantially reduces phase interpretability, and treated the 4 antennas within each recording as repeated views rather than independent observations. For the complex signal at antenna a, $z_a[n]$, we measured adjacent-sample phase increments,

$$\Delta\phi_a[n] = \angle(z_a[n+1]z_a[n]^*)\,, \tag{31}$$

excluding comparisons whose magnitude was below 10% of that antenna's RMS amplitude. We summarize local phase regularity by the circular concentration of $\Delta\phi_a[n]$ within 16 normalized windows and summarize variation across each recording by the circular spread of the window-level mean increments. Because generated and reference recordings are independent draws, the evaluation compares their distributions rather than requiring pointwise phase agreement.

Figure 15 shows that ALF closely follows the reference phase behavior for recording lengths from 2k through 32k samples. In particular, the median phase-rate drift remains small and does not increase with recording length: it is 0.066 rad/sample at 2k samples and 0.047 rad/sample at 32k samples for generated recordings, compared with 0.071 and 0.048 rad/sample for the reference recordings. Local phase-increment concentration is similarly stable across recording lengths. The waveform-class breakdown in Figure 16 shows close agreement for most waveform families. CW and pulsed CW are the most stringent cases because their definitions impose unusually regular phase evolution: CW shows the clearest long-range gap, while pulsed CW has modest differences in local concentration and drift. These exceptions do not dominate the aggregate result. Overall, the length-stratified and waveform-class-stratified results support that ALF produces phase-consistent recordings without systematic degradation at longer recording lengths.

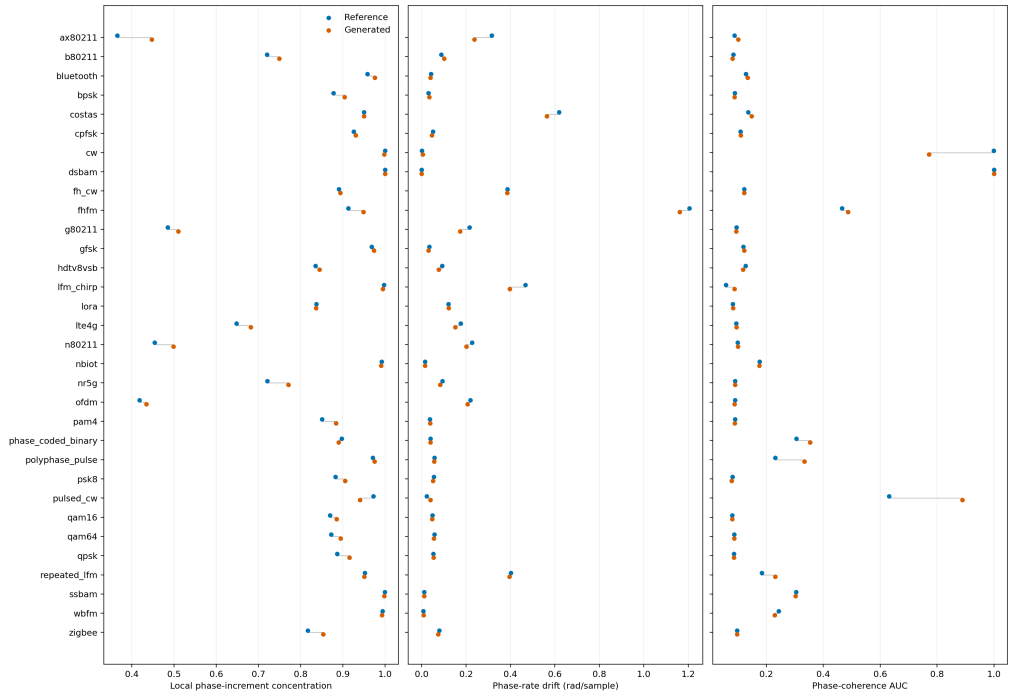


Figure 16: Within-record phase behavior by waveform family. Each point is the median across records; connecting lines pair generated and reference distributions for each waveform. The panels report local phase-increment concentration, phase-rate drift across normalized windows, and phase-only coherence averaged over normalized lags, where closer reference and generated dots means better performance.

E.10 Cross-Antenna Consistency

We evaluated whether ALF preserves array-level structure using the same 100K channel-held-out conditioning set as in Section E.9. We again retained the 71,673 recordings with stored mean per-antenna pre-combining SNR $>$ 10 dB and compared each generated 4-antenna recording with its corresponding reference recording under the same metadata. This test probes a property often omitted from RF-generation evaluations: realistic marginal antenna signals alone do not establish that a model has learned the shared timing and relative-phase structure of a multi-antenna recording.

For an adjacent antenna pair $(a, a+1)$, we estimate the relative delay from the peak of the overlap-normalized, demeaned complex cross-correlation,

$$\hat{\tau}_{a,a+1} = \arg \max_{\ell \in [-16,16]} \left| \frac{\sum_{n \in \mathcal{I}_\ell} \tilde{z}_a[n+\ell] \tilde{z}^*_{a+1}[n]}{\sqrt{\sum_{n \in \mathcal{I}_\ell} |\tilde{z}_a[n+\ell]|^2 \sum_{n \in \mathcal{I}_\ell} |\tilde{z}_{a+1}[n]|^2}} \right|, \tag{32}$$

where $\tilde{z}_a$ is the mean-removed complex signal at antenna a and $\mathcal{I}_\ell$ contains the valid sample indices at lag ℓ. We report the within-recording adjacent-pair spread,

$$\Delta_\tau = \max_{a \in \{1,2,3\}} \hat{\tau}_{a,a+1} - \min_{a \in \{1,2,3\}} \hat{\tau}_{a,a+1}. \tag{33}$$

A small Δ_τ indicates that all adjacent pairs share the same bulk timing relationship. We restrict this measure to recordings with identifiable correlation peaks, which includes approximately 89% of both generated and reference recordings.

To assess temporal relative-phase stability, we divide each recording into 16 normalized windows, estimate the phase of the complex adjacent-pair correlation in each window, unwrap the resulting phase trajectory, and fit its weighted slope $\hat{\beta}_{a,a+1}$. The reported recording-level phase drift is

$$D_\phi = \max_{a \in \{1,2,3\}} \frac{|\hat{\beta}_{a,a+1}|(N_{\text{samples}} - 1)}{2\pi}, \tag{34}$$

measured in cycles over the full recording. This metric tests for antenna-specific frequency offsets or time-varying relative phase; it does not require equal static phase across antennas, since spatial separation naturally produces antenna-dependent path phases.

Figure 17 shows that ALF preserves tightly shared adjacent-pair timing across all recording lengths: the median delay spread is zero samples in every length bin, with generated interquartile ranges no larger than one sample. Relative phase drift is also low and does not worsen for longer recordings. ALF has modestly larger drift than the reference distribution—from 0.023 cycles at 2k samples to 0.0055 cycles at 32k samples, versus 0.0105 and 0.0027 cycles for the reference recordings—but these values remain small over each complete recording. Thus, ALF learns the principal shared-timing and stable relative-phase behavior of the receive array.

The effective-rank results qualify this conclusion. ALF has a slightly lower spatial-covariance effective rank, particularly in the upper tail, which indicates modest compression of antenna-specific multipath diversity and is consistent with increased pairwise correlation. Thus, ALF does not perfectly reproduce every array statistic. Nevertheless, the timing and phase-drift results show that generated antenna streams remain coherently related over long recordings, which is the central array-level property evaluated here.

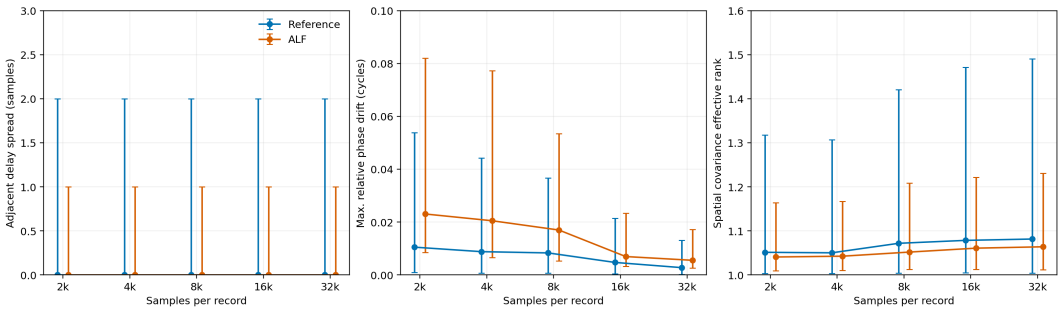


Figure 17: Cross-antenna consistency by recording length. Points show medians and bars show interquartile ranges. Adjacent delay spread measures agreement among the three adjacent-pair correlation-peak lags; relative phase drift measures the largest change in adjacent-pair correlation phase over a recording; and effective rank summarizes spatial-covariance diversity.

F AUTOREGRESSIVE POST-TRAINING

ALF is trained to sample a recording's P patch tokens jointly up to $\max(P) = 32$. This equates to a maximum generation length of 32,768 samples long. Although we have shown that these generation lengths are achieved with SOTA quality, additional downstream tasks like integrated sensing will require data that is orders of magnitude longer. Thus, we extend ALF to conditional suffix generation, $p(x_{k:P-1} \mid x_{0:k-1})$, through post-training a secondary mode. Not only can we demonstrate wireless signal generation of arbitrary length, but we can also exhibit custom label-driven distribution shifts (Figure 18).

Other diffusion models similarly extend inference with an autoregressive loop. Google's DiffusionGemma denoises a canvas, writes a copy to a context window, and then generates the next block with added conditioning (DiffusionGemma Team et al., 2026). ALF uses the same high-level windowed conditioning approach, but maintains continuity by explicitly using the last k patches and their clean latent representations as anchors to generate the following $P - k$ patches based on their labels. This method preserves the RF signal structure, resulting in smoother transitions.

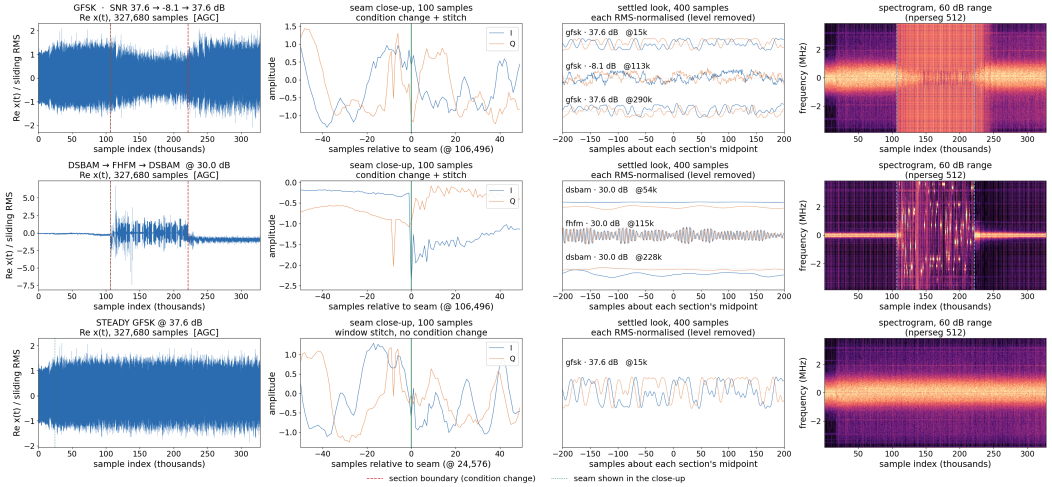


Figure 18: Autoregressive adaptation of AL generating 327,680-length samples. **Top row:** a pulsed CW signal transitions from 44.9 to −13.8dB in SNR and back (noise shift). **Middle row:** Bluetooth signal transitions to repeated LFM and back (waveform shift). **Bottom row:** steady pulsed CW signal at 44.9 dB provides a no-transition control.

We selected $P = 16, k = 8$ for post-trained ALF from a warm-start checkpoint seen in Figure 18. Training is split into two sections. **Mode 0** keeps the same anchored flow matching objective $\mathcal{L}_{\text{Stage 2}}$ described in section 5.2. **Mode 1** continuity training is applied to the first half of each length group, using independent noise times and prefix lengths. The sampled prefix is then assigned a zero velocity target, effectively masking its flow while still being used for mixing. The total objective combines flow matching over unmasked suffix tokens, a zero-velocity pin loss on prefix tokens, and an edge loss on the first token after the prefix.

$$\mathcal{L}_{\text{free}} = \frac{\sum_{b,i:\, m_{b,i}=0} \left\| \hat{\mathbf{v}}_{b,i} - \mathbf{v}^{*c}_{b,i} \right\|_2^2}{N_{\text{free}} D}, \qquad \mathcal{L}_{\text{pin}} = \frac{\sum_{b,i:\, m_{b,i}=1} \left\| \hat{\mathbf{v}}_{b,i} \right\|_2^2}{N_{\text{given}} D}, \tag{35}$$

$$\mathcal{L}_{\text{edge}} = \frac{1}{BD} \sum_b \left\| \hat{\mathbf{v}}_{b,k_b} - \mathbf{v}^{*c}_{b,k_b} \right\|_2^2, \qquad \hat{\mathbf{v}} = v_\theta(\mathbf{x}^c_t, t^c, c, \mathbf{m}). \tag{36}$$

$$\mathcal{L}_{\text{cont}} = \mathcal{L}_{\text{free}} + \lambda_{\text{pin}} \mathcal{L}_{\text{pin}} + \lambda_{\text{edge}} \mathcal{L}_{\text{edge}}, \qquad \lambda_{\text{pin}} = 0.1, \lambda_{\text{edge}} = 0.3 \tag{37}$$

The total post-training loss combines this objective with the original generation loss with unit weight,

$$\mathcal{L} = \mathcal{L}_{\text{Stage 2}} + \lambda_{\text{cont}} \mathcal{L}_{\text{cont}}, \qquad \lambda_{\text{cont}} = 1 \tag{38}$$

G Additional Figures and Tables

G.1 Distributional Metrics

Table 14: Comparison across PSD and spectral MSE metrics.

Length	Method	PSD MSE ↓	Spec MSE ↓
2,048	WaveStitch	158.08	379.50
	RF Diffusion†	925.4	1094.4
	Time Weaver	44.57	122.39
	ALF (ours)	**20.42**	**117.36**
4,096	WaveStitch	163.07	399.07
	Time Weaver	46.55	128.38
	ALF (ours)	**17.17**	**117.35**
8,192	WaveStitch	164.21	403.27
	Time Weaver	42.98	120.78
	ALF (ours)	**15.91**	**119.77**
16,384	WaveStitch	166.73	406.38
	Time Weaver	41.50	**114.90**
	ALF (ours)	**15.79**	119.00
32,768	WaveStitch	169.61	421.28
	Time Weaver	44.00	**112.91**
	ALF (ours)	**14.96**	122.18
Avg.	WaveStitch	164.3400	401.9000
	Time Weaver	43.9200	119.8720
	ALF (ours)	**16.8500**	**119.1320**
ALF Advantage	**vs. WaveStitch**	**89.75%**	**70.36%**
	vs. RF Diffusion†	**97.79%**	**89.28%**
	vs. Time Weaver	**61.63%**	**0.62%**

† RF Diffusion could not be trained for records of length greater than 2,048 due to memory limitations; longer generations were not considered. All other models are evaluated on 100k recordings.

Table 14 compares ALF against its competitors using more distributional metrics. Although downstream tasks are a better judgment of quality, we include these to provide a full view of ALF's performance (Baur et al. (2025)). The table shows we are significantly superior in PSD MSE, and overall better in Spectrogram MSE. However, this further validates our claim that distributional alignment is not fine enough of an evaluation metric for highly constrained representations of wireless signals.

G.2 Other Generation Quality Metrics

Table 15: Effect on quality metrics due to rescaling on energy-linked objectives

		Delay Spread ($k = 9$)		Indoor/Outdoor ($k = 2$)		Urban/Rural ($k = 2$)		LOS/NLOS ($k = 2$)	
Power mode	Method	TSTR	TRTS	TSTR	TRTS	TSTR	TRTS	TSTR	TRTS
Normalized	WaveStitch	0.11	0.11	0.50	0.49	0.50	0.51	0.49	0.53
	TimeWeaver	0.17	0.15	0.54	**0.64**	**0.57**	0.54	0.56	0.61
	ALF	**0.19**	**0.18**	**0.56**	0.64	0.54	**0.56**	**0.57**	**0.71**
Rescaled	WaveStitch	0.11	0.12	0.50	0.74	0.50	0.55	0.56	0.75
	TimeWeaver	0.20	0.16	0.74	0.73	**0.57**	0.56	0.84	0.76
	ALF	**0.22**	**0.20**	**0.75**	**0.74**	0.56	**0.57**	**0.85**	**0.79**
Rescaling Gain (Δ)	WaveStitch	0.00	0.01	-0.01	0.24	0.00	0.04	0.06	0.22
	TimeWeaver	0.04	0.01	0.20	0.09	-0.00	0.02	0.28	0.15
	ALF	0.04	0.02	0.19	0.10	0.02	0.00	0.28	0.08
Above Chance (Δ_{chance})	WaveStitch	0.00	0.01	0.00	**0.24**	0.00	0.05	0.06	0.25
	TimeWeaver	0.09	0.05	0.24	0.23	**0.07**	0.06	0.34	0.26
	ALF	**0.11**	**0.09**	**0.25**	**0.24**	0.06	**0.07**	**0.35**	**0.29**

Note: k = number of classes; pooled = mean over lengths 2,048–32,768. Δ = Rescaled minus Normalized. Δ_{chance} = Rescaled minus chance accuracy, where chance is $1/k$. LOS = Line-of-Sight; NLOS = No-Line-of-Sight.

Table 15 shows the rest of the classification objectives unable to fit into the main paper (Table 2, § 6). These tasks show lower utility performance than Time Weaver; however, this is due to their energy-based nature. Both models output normalized data that lacks attenuation, which makes scene classifications (Indoor/Outdoor, LOS/NLOS) far easier to detect than shape alone. After rescaling each recording to match its ground-truth reference from each respective class, we see classification performance equalize across WaveStitch, Time Weaver, and ALF for Indoor/Outdoor and LOS/NLOS tasks. The other two objectives have accuracies barely over their chance probabilities. These noise readings indicate these as more difficult representations.