ANCHOR-PRESERVING COARSE-TO-FINE DIFFUSION FOR PROBABILISTIC TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Probabilistic time series forecasts should represent uncertainty at both individual time steps and aggregated temporal resolutions. Reliable uncertainty at individual steps, however, does not necessarily imply reliable aggregate uncertainty. Coarse-to-fine generation can use deterministic coarse forecasts to guide refinement. Yet fixing every sample's aggregate to the same forecast eliminates aggregate uncertainty, while conditioning alone does not guarantee that the aggregate ensemble mean matches it. Preliminary experiments on two datasets show that combining a point forecast with centered stochastic deviations reduces CRPS. With fixed final coarse scenarios, controlled comparisons provide limited support for centering before detail generation. We propose an anchor-preserving coarse-to-fine diffusion framework that constrains the coarse ensemble mean rather than individual aggregate outcomes. Residual diffusion generates stochastic coarse scenarios, which are centered around a deterministic anchor before conditional detail generation. A conditional detail diffusion model then refines each scenario using projected detail with zero block averages, so refinement cannot alter its coarse aggregate. The resulting forecasts satisfy sample-wise coherence and coarse ensemble-mean preservation up to numerical precision while retaining stochastic aggregate variation. The proposed method achieves the lowest mean CRPS in 15 of 24 dataset–horizon settings among methods with complete three-seed results, but substantially underperforms on Electricity and Traffic. Matched-mean evaluation of shared forecast ensembles further shows that fine-resolution gains need not translate into better aggregate forecasts.

1 INTRODUCTION

Probabilistic time series forecasting models a distribution over future trajectories rather than predicting a single sequence. A forecast can represent uncertainty well at individual time steps while misrepresenting uncertainty in averages or totals over future intervals. Aggregate uncertainty depends on temporal dependence as well as the marginal distributions at individual steps. When forecasts are constructed at multiple temporal resolutions, they should also remain consistent under aggregation (Rangapuram et al., 2021; 2023). Probabilistic reconciliation can enforce this consistency while changing forecast means and variances (Zambon et al., 2024). Preserving a chosen coarse forecast as the ensemble mean is therefore a separate requirement.

Multi-resolution forecasting captures broad trends and local variations through representations at different scales (Wang et al., 2024). In probabilistic forecasting, mr-Diff implements this principle by generating coarse trends and progressively refining them with conditional diffusion models (Shen et al., 2024). A complementary line of work combines deterministic forecasts with stochastic generation (Li et al., 2024; 2025; Ye et al., 2025; Lai et al., 2025). These ideas motivate coarse-to-fine generation guided by a deterministic coarse forecast.

We consider settings where a separately trained coarse predictor provides a useful reference location. The question is how to preserve this reference without treating it as the only possible aggregate outcome. For an interval-average forecast of 100, scenarios averaging 80 and 120 preserve the ensemble mean while retaining aggregate variation. Requiring each scenario to average 100 removes this variation; conditioning on 100 alone does not ensure the desired ensemble mean. We therefore

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.

distinguish the ensemble-level requirement of preserving the coarse mean from the sample-level requirement of preserving each generated coarse scenario during refinement.

To examine the role of the coarse anchor, we perform a preliminary experiment. Let Q_s be the aggregated D^{3}U samples and $\bar{Q}$ their ensemble mean. Using a deterministic TimeMixer forecast $\widehat{C}$, we construct $Q'_s = \widehat{C} + (Q_s - \bar{Q})$. This replaces the ensemble center while preserving the centered sample deviations. As shown in Table 1, the point-only forecast has lower RMSE but higher CRPS than the original D^{3}U on both datasets. The recentered ensemble matches the point forecast's RMSE by construction and achieves lower CRPS than either alternative. This motivates preserving a useful coarse forecast anchor without discarding stochastic aggregate variation.

Table 1: Effect of coarse recentering on ETTh1 and ETTh2 validation forecasts at horizon 96 and block size 4, using 100 samples per window. Bold marks the lowest mean per dataset and metric.

	ETTh1		ETTh2	
Condition	RMSE ↓	CRPS ↓	RMSE ↓	CRPS ↓
Original D^{3}U	0.5558 ± 0.014	0.2745 ± 0.0088	0.5042 ± 0.0023	0.2680 ± 0.0041
Point only (TimeMixer)	$\mathbf{0.5095 \pm 0.0026}$	0.3319 ± 0.00084	$\mathbf{0.4697 \pm 0.0034}$	0.3084 ± 0.0027
Point-anchored D^{3}U	$\mathbf{0.5095 \pm 0.0026}$	$\mathbf{0.2457 \pm 0.00087}$	$\mathbf{0.4697 \pm 0.0034}$	$\mathbf{0.2481 \pm 0.0032}$

Post-hoc mean correction can also match the coarse ensemble mean, but a common blockwise shift leaves existing within-block detail unchanged. We propose an *anchor-preserving coarse-to-fine diffusion framework* that instead centers coarse scenarios before generating their conditional detail. A residual diffusion model supplies stochastic deviations from a deterministic coarse anchor. We center these deviations before forming coarse scenarios. A second diffusion model generates detail conditioned on each centered scenario, with zero block averages enforced throughout sampling. The anchor therefore constrains the ensemble mean rather than every aggregate outcome, and refinement cannot change the coarse scenario used to generate its detail. Both constraints hold up to numerical precision, independently of denoising accuracy. We build on ensemble recentering (Keppenne, 2014) and block-average-preserving reconstruction (Harder et al., 2023; Salatiello et al., 2024).

We evaluate the same forecast ensembles across temporal resolutions and use matched-mean controls to assess aggregate forecasts after aligning their coarse ensemble means. In the experiments, our model achieves the lowest mean fine-resolution CRPS in 15 of 24 dataset–horizon settings. On ETTh1 and Weather at horizon 96, matched-mean comparisons show that fine-resolution gains need not translate into better aggregate forecasts. The fixed configuration performs poorly on Electricity and Traffic, highlighting the importance of coarse-distribution quality.

The contributions of this work are summarized as follows.

- We distinguish ensemble-level anchor preservation from sample-wise coherence, allowing a coarse point forecast to constrain the ensemble mean without fixing individual aggregate outcomes.
- We introduce a coarse-to-fine diffusion sampler that centers stochastic coarse scenarios before conditional detail generation and enforces both requirements up to numerical precision.
- Our experiments suggest that centering order can affect fine-resolution quality even when final coarse scenarios are identical, and that fine-resolution gains need not extend to aggregate forecasts under matched coarse means.

For reproducibility, source code will be released upon acceptance.

2 Background

2.1 Point-guided and residual probabilistic generation

Diffusion models have been adapted to forecasting through both autoregressive and non-autoregressive generation. TimeGrad generates future observations autoregressively using condi-

tional denoising diffusion (Rasul et al., 2021), whereas TimeDiff develops a non-autoregressive forecasting model with dedicated conditioning mechanisms (Shen & Kwok, 2023).

A complementary direction combines point prediction with probabilistic generation. CARD couples a pretrained conditional mean estimator with a conditional diffusion model (Han et al., 2022). In time series forecasting, TMDM (Li et al., 2024) incorporates transformer predictions into the diffusion process, while D^3U (Li et al., 2025) separates deterministic prediction from uncertainty modeling. Preserving point-prediction accuracy while learning uncertainty is also the motivation behind faithful heteroscedastic regression (Stirn et al., 2023). NsDiff (Ye et al., 2025) incorporates estimated conditional means and variances into a non-stationary diffusion process. We focus on preserving a prescribed coarse ensemble mean. Each fine-resolution sample must also aggregate to the coarse scenario used to generate it.

2.2 Multiscale and coarse-to-fine modeling

Multiscale forecasting models exploit temporal structure at different resolutions. Scaleformer iteratively refines forecasts across scales (Shabani et al., 2023), while TimeMixer (Wang et al., 2024) combines decomposed temporal patterns through multiscale mixing. These approaches provide a basis for separating broad temporal dynamics from finer variations.

Multiscale structure can also organize probabilistic generation. Wavelet score-based generative modeling (Guth et al., 2022) factorizes a data distribution through conditional generation across wavelet scales. For time series forecasting, mr-Diff (Shen et al., 2024) generates coarse trends and progressively refines them using finer conditional diffusion models. MG-TSD (Fan et al., 2024) also exploits temporal granularity by using coarse-grained targets to guide diffusion learning. Our construction instead uses an explicit aggregation operator to constrain the final coarse–fine relationship, together with a separate constraint on the realized coarse ensemble mean.

2.3 Coherent forecasting and prescribed forecast preservation

Temporal hierarchies reconcile forecasts across temporal aggregation levels (Athanasopoulos et al., 2017). Coherent probabilistic forecasting extends consistency to predictive distributions over hierarchical and temporally aggregated series (Rangapuram et al., 2021; 2023). Probabilistic reconciliation also connects these structural constraints with distributional evaluation and score-based optimization (Panagiotelis et al., 2023).

A related requirement is to preserve selected forecasts during reconciliation. Immutable forecast reconciliation keeps a prescribed subset of point forecasts unchanged while adjusting the remaining forecasts to satisfy coherence (Zhang et al., 2023). Our setting instead specifies the arithmetic mean of a finite coarse forecast ensemble while retaining variation among its members. A completed fine-resolution ensemble can also be translated to satisfy this mean constraint. We therefore distinguish the constraint itself from our generation order, which centers coarse scenarios before producing their conditional detail.

3 Proposed Method

Given history $X \in \mathbb{R}^{L\times D}$, we generate an ensemble $\{Y_s\}_{s=1}^{S}$ with $Y_s \in \mathbb{R}^{H\times D}$. The construction assigns separate roles to the coarse anchor, coarse variation, and within-block detail. We first state the representation and its guarantees, then describe how the components are learned and sampled. The batch dimension is omitted throughout.

3.1 Coarse and Detail Representation

Let $H = bK$, where b is the architectural aggregation factor. Define block averaging $A : \mathbb{R}^{H\times D} \to \mathbb{R}^{K\times D}$ and repeat upsampling $U : \mathbb{R}^{K\times D} \to \mathbb{R}^{H\times D}$ by:

$$\begin{aligned} \big[AY\big]_{k,d} &= \frac{1}{b}\sum_{j=1}^{b} Y_{(k-1)b+j,d}, \\ \big[UC\big]_{(k-1)b+j,d} &= C_{k,d}, \end{aligned} \tag{1}$$

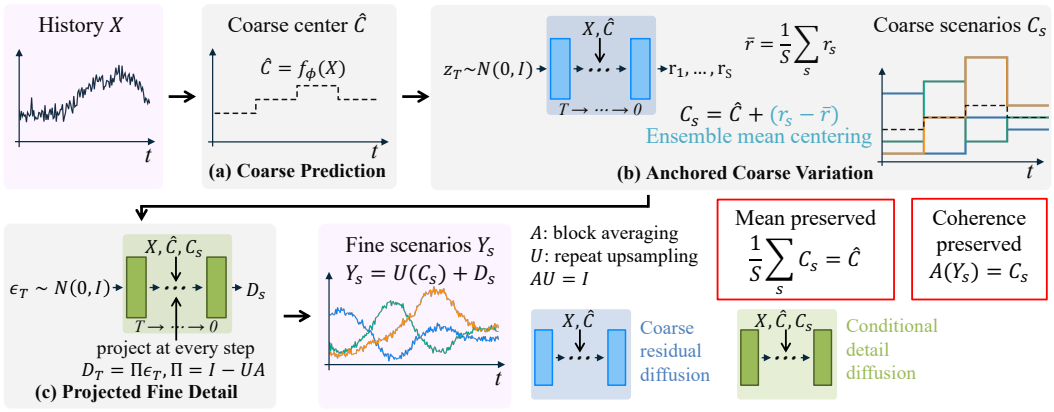

Figure 1: Anchor-preserving coarse-to-fine generation. (a) A point predictor supplies the coarse anchor. (b) Centered residual samples yield coarse scenarios whose ensemble mean equals the anchor. (c) Scenario-conditioned, projected detail preserves each scenario's block averages. Matching colors pair coarse and fine scenarios, and black dashed lines indicate the anchor.

for $k = 1, \ldots, K$, $j = 1, \ldots, b$, and $d = 1, \ldots, D$. These operators act independently on each channel and satisfy $AU = I_K$. The detail projector and target decomposition are:

$$\Pi = I_H - UA, \qquad A\Pi = 0, \qquad \Pi^2 = \Pi, \tag{2}$$

$$C^\star = AY, \qquad D_0 = \Pi Y, \qquad Y = UC^\star + D_0. \tag{3}$$

Thus, $C^\star$ contains the block averages and $D_0 \in \ker A$ contains variation that leaves those averages unchanged. This decomposition is related to the range–null-space construction used by DDNM for measurement-consistent image restoration (Wang et al., 2023). Closely related mean-preserving reconstructions appear in deterministic temporal hierarchical forecasting with block-centered decoder outputs (Salatiello et al., 2024). We use the same block-centering principle with stochastic coarse scenarios and additionally constrain their realized ensemble mean.

Let $\widehat{C}$ be the predicted coarse anchor, r_s a generated coarse residual, and $\bar{r} = \frac{1}{S}\sum_s r_s$. The complete forecast has the form:

$$Y_s = U\Bigg(\underbrace{\widehat{C}}_{\text{coarse anchor}} + \underbrace{r_s - \bar{r}}_{\text{coarse variation}}\Bigg) + \underbrace{D_s}_{\text{within-block detail}}, \qquad AD_s = 0. \tag{4}$$

Writing $C_s = \widehat{C} + r_s - \bar{r}$, this construction gives:

$$AY_s = C_s, \qquad \frac{1}{S}\sum_{s=1}^{S} AY_s = \widehat{C}, \qquad C_s - C_{s'} = r_s - r_{s'}. \tag{5}$$

Individual aggregate outcomes may therefore vary, their ensemble mean is fixed to the anchor, and detail does not change any sample's block averages. These identities hold for the chosen aggregation operator independently of the accuracy of the learned models. Figure 1 summarizes the generation order.

3.2 Coarse Prediction

A deterministic predictor maps the history to a coarse anchor,

$$\widehat{C} = f_\phi(X) \in \mathbb{R}^{K \times D}. \tag{6}$$

We use a TimeMixer (Wang et al., 2024) backbone with a coarse-resolution output head and train it against the aggregated future using mean absolute error (MAE). Among all predictors, expected

absolute error is minimized by conditional medians, which need not equal conditional means (Gneiting, 2011). We use the fitted prediction as a prescribed coarse reference location. Enforcing this prediction as the ensemble mean is a modeling choice, not a claim that MAE estimates the true conditional mean.

Both stochastic components are trained using cross-fitted coarse predictions $\widetilde{C}$ from predictors fitted without the corresponding held-out training windows. This provides held-out prediction residuals while keeping the location predictor separate from the diffusion models.

3.3 Anchored Coarse Variation

For each training example, form $R = C^\star - \widetilde{C}$ and normalize it using entrywise statistics fitted only to training out-of-fold residuals:

$$Z_0 = (R - m) \oslash v, \qquad v_{k,d} \geq \varepsilon > 0. \tag{7}$$

Here m and v are the fitted mean and standard deviation, and $\oslash$ and $\odot$ denote elementwise division and multiplication. We use the Gaussian forward process and noise-prediction objective of DDPM (Ho et al., 2020) in normalized residual coordinates. For $t = 1, \ldots, T$, let $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{u=1}^{t} \alpha_u$, and $\bar{\alpha}_0 = 1$. The forward process is:

$$Z_t = \sqrt{\bar{\alpha}_t}\, Z_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{8}$$

We use the shifted input $W_t = \widetilde{C} + Z_t$ for the coarse denoiser. This shift is a network-input parameterization; the diffusion state and reverse updates remain in normalized residual coordinates. A conditional denoiser predicts the injected noise with the objective:

$$\mathcal{L}_{\text{coarse}} = \mathbb{E}_{(X,Y),t,\epsilon}\left[\frac{1}{KD}\|\epsilon_\psi(W_t, t; X, \widetilde{C}) - \epsilon\|_F^2\right], \tag{9}$$

where t is sampled uniformly. The denoiser conditions on history and the coarse anchor.

At inference, $\widetilde{C}$ is replaced by $\widehat{C} = f_\phi(X)$. Reverse trajectories start from $Z_T^{(s)} \sim \mathcal{N}(0, I)$ and produce $r_s = m + v \odot Z_0^{(s)}$. We apply ensemble recentering, a common-shift adjustment used in ensemble data assimilation (Keppenne, 2014):

$$\bar{r} = \frac{1}{S}\sum_{s=1}^{S} r_s, \qquad C_s = \widehat{C} + r_s - \bar{r}. \tag{10}$$

Here, the target center is the deterministic coarse anchor, and recentering is performed before scenario-conditioned detail generation. The common shift preserves the coarse ensemble's empirical spread. Fine-resolution spread can still change because detail is generated conditional on the centered scenarios.

3.4 Projected Fine Detail

Diffusion processes restricted to zero-sum linear subspaces have been used in molecular generation (Hoogeboom et al., 2022), and projection-based denoising has also been studied for constrained time-series generation (Narasimhan et al., 2026). We restrict both the detail forward process and the reverse sampler to the aggregation null space. At inference, the centered coarse scenario conditions detail generation and is also used in the final reconstruction. The detail forward process remains in the aggregation null space:

$$D_t = \sqrt{\bar{\alpha}_t}\, D_0 + \sqrt{1 - \bar{\alpha}_t}\, \Pi\xi, \qquad \xi \sim \mathcal{N}(0, I). \tag{11}$$

Since $AD_0 = A\Pi\xi = 0$, every noised state satisfies $AD_t = 0$. For block averaging and repeat upsampling, Π is an orthogonal projector; projected noise has temporal covariance Π.

During training, the detail denoiser receives the true coarse target $C^\star$ and the separate out-of-fold location $\widetilde{C}$:

$$\begin{aligned} V_{\theta,t} &= h_\theta(UC^\star + D_t, t; X, UC^\star, U\widetilde{C}), \\ \widehat{D}_{0|t} &= \Pi V_{\theta,t}. \end{aligned} \tag{12}$$

It predicts clean detail using:

$$\mathcal{L}_{\text{detail}} = \mathbb{E}_{(X,Y),t,\xi}\left[\frac{\|\widehat{D}_{0|t} - D_0\|_F^2}{HD} + \lambda_{\text{leak}}\frac{\|AV_{\theta,t}\|_F^2}{KD}\right]. \tag{13}$$

The leakage penalty discourages coarse content in the raw network output. Exact coherence is enforced by projection.

At inference, the coarse target is replaced by C_s and the location condition by $\widehat{C}$. Starting from $D_T^{(s)} = \Pi\xi_T^{(s)}$, for selected reverse times $t > \tau \geq 0$ we use:

$$\begin{aligned}
\widehat{D}_{0|t}^{(s)} &= \Pi h_\theta(UC_s + D_t^{(s)}, t; X, UC_s, U\widehat{C}),\\
\widehat{\xi}_t^{(s)} &= \Pi\left(\frac{D_t^{(s)} - \sqrt{\bar{\alpha}_t}\,\widehat{D}_{0|t}^{(s)}}{\sqrt{1-\bar{\alpha}_t}}\right),\\
D_\tau^{(s)} &= \Pi\left(\sqrt{\bar{\alpha}_\tau}\,\widehat{D}_{0|t}^{(s)} + \sqrt{1-\bar{\alpha}_\tau}\,\widehat{\xi}_t^{(s)}\right).
\end{aligned} \tag{14}$$

We use deterministic DDIM sampling (Song et al., 2021) with the same centered scenario C_s at every step. Projection at initialization, prediction, and each reverse update keeps the detail in $\ker(A)$. The final detail $D_s = D_0^{(s)}$ is combined with the coarse scenario as $Y_s = UC_s + D_s$.

3.5 Training and Forecast Generation

The location predictor and the two diffusion models are trained separately. Five-fold purged blocked cross-fitting supplies out-of-fold locations: training windows whose history or future overlaps a held-out window are excluded from the corresponding fold fit. A final location predictor is fitted on the full training split for inference. Residual normalization statistics are fitted to the training out-of-fold residuals and frozen. The diffusion models optimize Eqs. (9) and (13) with the location predictions held fixed. Observed future targets are used for supervised training and validation, not as inputs to test-time forecast generation.

For each history, we generate and center all $S = 100$ coarse scenarios before sampling their paired details. The ensemble mean provides a point forecast, $A\bar{Y} = \widehat{C}$. Although $A\bar{D} = 0$, the fine-step entries of $\bar{D}$ need not vanish; coarse mean preservation does not force $\bar{Y} = U\widehat{C}$. Empirical quantiles provide prediction intervals. The algebraic identities hold up to floating-point error, whereas accuracy and interval coverage are empirical properties.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets

We evaluate on six multivariate time series datasets: ETTh1, ETTh2, ETTm1, Weather, Electricity, and Traffic. We use context length $L = 336$, forecasting horizons $H \in \{96, 192, 336, 720\}$, and 70/10/20 train/validation/test splits. Channel-wise standardization is fitted only on the training split.

4.1.2 Metrics

We assess probabilistic forecasts using the continuous ranked probability score (CRPS) (Gneiting & Raftery, 2007) and central 90% empirical interval coverage. For CRPS, targets and samples are standardized before aggregation. Lower CRPS is better, and coverage closer to 0.90 is preferred. We use q to denote the evaluation aggregation length, with q corresponding to the original temporal resolution.

4.1.3 Baselines

We compare with three diffusion-based forecasting methods, D^3U (Li et al., 2025), NsDiff (Ye et al., 2025), and mr-Diff (Shen et al., 2024), and a simple probabilistic baseline, TM+G. TM+G pairs a

full-resolution TimeMixer predictor (Wang et al., 2024) with independent Gaussian residuals whose per-channel, per-lead mean and variance are fitted on the training windows; it does not use the proposed model's coarse anchor. All baselines generate $S = 100$ samples per forecast.

For numerical stability, we modify NsDiff in one place: the log of the de-stationary attention temperature τ is clamped to $[-5, 5]$, because the unbounded τ overflowed in FP32 on some batches. mr-Diff is our reimplementation, with moving-average kernels $\{5, 25, 51, 201\}$, 100 diffusion steps for training and sampling, and checkpoint selection by validation MAE. Our mr-Diff reimplementation produces strongly underdispersed ensembles under the evaluation protocol used here.

4.1.4 Implementation Details

Our fixed configuration uses aggregation factor $b = 4$, a TimeMixer coarse-location predictor, and coarse and detail denoisers with hidden widths 48 and 64, respectively. Both diffusion components use 100 training timesteps and 20 sampling steps. We generate $S = 100$ samples per forecast and report results over training seeds 1–3, excluding incomplete seed groups from ranking. Residual preparation uses five-fold purged blocked cross-fitting. The main benchmark and multi-resolution comparison reuse the corresponding fitted models without scalar spread adjustment.

4.2 Results

Table 2: Fine-resolution standardized CRPS (↓), averaged over three training seeds using 100 samples per test window. mr-Diff denotes our reimplementation. Bold marks the lowest complete mean.

Dataset	H	TM+G	D^{3}U	NsDiff	mr-Diff	Ours
ETTh1	96	0.3347 ± 0.0016	0.3504 ± 0.0090	0.3846 ± 0.0056	0.5052 ± 0.0207	$\mathbf{0.3261 \pm 0.0012}$
	192	0.3642 ± 0.0005	0.3682 ± 0.0010	0.4303 ± 0.0172	0.5267 ± 0.0106	$\mathbf{0.3531 \pm 0.0028}$
	336	0.3916 ± 0.0023	0.3984 ± 0.0032	0.4608 ± 0.0075	0.5804 ± 0.0154	$\mathbf{0.3813 \pm 0.0046}$
	720	0.4531 ± 0.0010	0.4866 ± 0.0135	0.4866 ± 0.0184	0.6833 ± 0.0065	$\mathbf{0.4442 \pm 0.0025}$
ETTh2	96	0.2236 ± 0.0024	0.2429 ± 0.0068	0.2567 ± 0.0073	0.3549 ± 0.0056	$\mathbf{0.2076 \pm 0.0010}$
	192	0.2549 ± 0.0012	0.2666 ± 0.0089	0.2882 ± 0.0203	0.3786 ± 0.0032	$\mathbf{0.2348 \pm 0.0010}$
	336	0.2828 ± 0.0017	0.2806 ± 0.0047	0.2937 ± 0.0159	0.3991 ± 0.0042	$\mathbf{0.2585 \pm 0.0072}$
	720	0.3405 ± 0.0008	$\mathbf{0.3047 \pm 0.0063}$	0.3401 ± 0.0397	0.4516 ± 0.0237	0.3092 ± 0.0137
ETTm1	96	0.2763 ± 0.0006	0.2892 ± 0.0034	0.3676 ± 0.0238	2.1888 ± 2.5415	$\mathbf{0.2701 \pm 0.0019}$
	192	0.3062 ± 0.0016	0.3019 ± 0.0020	0.3653 ± 0.0147	1.0658 ± 0.2903	$\mathbf{0.2940 \pm 0.0008}$
	336	0.3330 ± 0.0014	0.3272 ± 0.0035	0.3919 ± 0.0064	0.6995 ± 0.0998	$\mathbf{0.3209 \pm 0.0014}$
	720	0.3679 ± 0.0012	0.3551 ± 0.0043	0.4136 ± 0.0125	0.8565 ± 0.0596	$\mathbf{0.3515 \pm 0.0008}$
Weather	96	0.2131 ± 0.0010	0.2029 ± 0.0027	0.2562 ± 0.0013	0.3562 ± 0.0142	$\mathbf{0.1896 \pm 0.0020}$
	192	0.2473 ± 0.0004	0.2420 ± 0.0072	0.2825 ± 0.0099	0.4279 ± 0.0298	$\mathbf{0.2215 \pm 0.0017}$
	336	0.2788 ± 0.0005	0.2726 ± 0.0033	0.3195 ± 0.0064	0.4593 ± 0.0282	$\mathbf{0.2542 \pm 0.0022}$
	720	0.3224 ± 0.0011	0.3101 ± 0.0036	0.3685 ± 0.0074	0.4652 ± 0.0009	$\mathbf{0.3002 \pm 0.0034}$
Electricity	96	$\mathbf{0.1725 \pm 0.0003}$	0.1791 ± 0.0010	0.2871 ± 0.0026	0.3806 ± 0.0123	0.8667 ± 0.0023
	192	$\mathbf{0.1864 \pm 0.0003}$	0.1908 ± 0.0009	0.2874 ± 0.0062	0.3542 ± 0.0135	0.9462 ± 0.0022
	336	$\mathbf{0.1999 \pm 0.0005}$	0.2022 ± 0.0015	0.2883 ± 0.0029	0.3653 ± 0.0045	1.0336 ± 0.0026
	720	$\mathbf{0.2229 \pm 0.0005}$	0.2238 ± 0.0003	0.2985 ± 0.0020	0.3750 ± 0.0045	1.1796 ± 0.0025
Traffic	96	0.2276 ± 0.0022	$\mathbf{0.2093 \pm 0.0023}$	0.4710 ± 0.0432	0.4381 ± 0.0430	1.1141 ± 0.0011
	192	0.2391 ± 0.0027	$\mathbf{0.2144 \pm 0.0017}$	0.5056 ± 0.1106	0.4073 ± 0.0115	1.1490 ± 0.0025
	336	0.2470 ± 0.0022	$\mathbf{0.2178 \pm 0.0013}$	0.4412 ± 0.0031	0.4946 ± 0.0669	1.1812 ± 0.0042
	720	0.2581 ± 0.0021	$\mathbf{0.2348 \pm 0.0057}$	0.5895 ± 0.0009	0.5505 ± 0.0627	1.2417 ± 0.0027

Table 2 shows our model achieves the lowest mean fine-resolution CRPS in 15 of 24 settings. These include most horizons on ETTh1, ETTm1, ETTh2, and Weather. However, the fixed configuration substantially underperforms on Electricity and Traffic.

5 Controlled Comparisons and Diagnostics

We first test whether centering order affects fine-resolution forecasts when the final coarse scenarios are held fixed. We then evaluate aggregate uncertainty after matching coarse ensemble means across methods.

5.1 Centering Before or After Detail Generation

We compare centering before and after detail generation while keeping the final coarse scenarios fixed. We use the existing checkpoints on ETTh1 and Weather with $H = 96$, $b = 4$, $S = 100$, and training seeds 1–3. Within each dataset and training seed, both variants use the same coarse residual samples and detail-sampling noise. No additional training or tuning is performed.

In pre-centering, detail generation is conditioned on the centered coarse scenario $C_s = \widehat{C} + r_s - \bar{r}$, where $\bar{r}$ is the ensemble mean of the coarse residuals. In post-centering, detail is generated conditioned on the uncentered coarse scenario $\widehat{C} + r_s$. We then apply the common shift $-U\bar{r}$ to the completed fine-resolution trajectories. Projection is retained throughout detail sampling in both variants.

Table 3: Pre-centering versus post-centering CRPS at $q = 1$. ΔCRPS represents $\mathrm{CRPS}_{\mathrm{post}} - \mathrm{CRPS}_{\mathrm{pre}}$. The two-sided paired t-test is applied to the three training-seed mean differences, equivalently testing $\Delta\mathrm{CRPS} = 0$ with $df = 2$.

Dataset	$\mathrm{CRPS}_{\mathrm{pre}}$	$\mathrm{CRPS}_{\mathrm{post}}$	ΔCRPS	$t(2)$	p
ETTh1	0.3261 ± 0.0012	0.3285 ± 0.0016	$+0.0024 \pm 0.0007$	5.5671	0.0308
Weather	0.1896 ± 0.0020	0.1903 ± 0.0023	$+0.0007 \pm 0.0003$	4.0798	0.0552

As shown in Table 3, pre-centering yields lower CRPS in all three seeds on both datasets, but only ETTh1 meets the unadjusted 5% significance threshold. The modest differences provide limited support for generating detail from the final centered coarse scenarios: fine-resolution quality can change even when aggregate forecasts are identical.

5.2 Aggregate Uncertainty under Matched Coarse Means

We evaluate ETTh1 and Weather at $H = 96$ by averaging forecasts and targets over aligned, nonoverlapping blocks of length $q \in \{1, 4, 12, 24\}$. For each model and input, the same $S = 100$ saved trajectories are used at every resolution.

For each baseline, D^{3}U or mr-Diff, we use the same coarse anchor $\widehat{C}$ as our model and apply

$$Y_s^{\mathrm{MC}} = Y_s^{\mathrm{base}} + U\left(\widehat{C} - A\bar{Y}^{\mathrm{base}}\right), \qquad \bar{Y}^{\mathrm{base}} = \frac{1}{S}\sum_{s=1}^{S} Y_s^{\mathrm{base}}, \tag{15}$$

where A and U use $b = 4$. This correction preserves sample deviations and empirical interval widths while matching ensemble means at $q = 4, 12, 24$. It does not generally match pointwise means at $q = 1$.

Table 4: Multi-resolution evaluation, using the same 100-sample fine-resolution ensemble at every aggregation length q. Bold marks the lowest mean CRPS.

Dataset	Method	CRPS ↓			
		$q = 1$	$q = 4$	$q = 12$	$q = 24$
ETTh1	mr-Diff	0.5052 ± 0.0207	0.4607 ± 0.0214	0.4002 ± 0.0116	0.3463 ± 0.0069
	Mean-corrected mr-Diff	0.4449 ± 0.0051	0.3894 ± 0.0035	0.3375 ± 0.0026	0.2988 ± 0.0021
	D^{3}U	0.3504 ± 0.0090	0.3155 ± 0.0079	0.2740 ± 0.0042	0.2453 ± 0.0046
	Mean-corrected D^{3}U	0.3280 ± 0.0034	0.2912 ± 0.0014	0.2518 ± 0.0011	0.2227 ± 0.0014
	Ours	$\mathbf{0.3261 \pm 0.0012}$	$\mathbf{0.2904 \pm 0.0011}$	$\mathbf{0.2511 \pm 0.0007}$	$\mathbf{0.2202 \pm 0.0005}$
Weather	mr-Diff	0.3562 ± 0.0142	0.3454 ± 0.0133	0.3312 ± 0.0112	0.3158 ± 0.0088
	Mean-corrected mr-Diff	0.2485 ± 0.0028	0.2264 ± 0.0019	0.2185 ± 0.0020	0.2109 ± 0.0020
	D^{3}U	0.2029 ± 0.0027	0.1931 ± 0.0030	0.1852 ± 0.0035	0.1784 ± 0.0035
	Mean-corrected D^{3}U	0.1932 ± 0.0016	$\mathbf{0.1825 \pm 0.0021}$	$\mathbf{0.1750 \pm 0.0024}$	$\mathbf{0.1682 \pm 0.0025}$
	Ours	$\mathbf{0.1896 \pm 0.0020}$	0.1839 ± 0.0020	0.1763 ± 0.0017	0.1688 ± 0.0013

Table 4 shows that mean correction reduces mean CRPS for both baselines at every evaluated resolution. This correction substantially narrows the CRPS gap between D^{3}U and our model. At the aggregate resolutions, our model has slightly lower mean CRPS on ETTh1, whereas mean-corrected D^{3}U

has slightly lower mean CRPS on Weather despite its higher fine-resolution CRPS. These observed rankings illustrate that fine-resolution improvements need not carry over to aggregate forecasts, even after matching aggregate ensemble means. Since projected detail cancels at $q = 4, 12, 24$, these aggregate scores assess coarse predictive distributions rather than the benefit of scenario-conditioned detail.

5.3 Effects of Centering and Coarse Stochasticity

Table 5 separates coarse-distribution error from changes in the forecast center. At $q = 4$, projected detail cancels, and the centered stochastic and deterministic variants share the same ensemble mean. Removing coarse variation sharply reduces CRPS on Electricity and Traffic, whereas centering alone has a much smaller effect. Together with coverage above 0.999, this identifies the generated coarse deviations as a major source of error. The deterministic variant still has zero aggregate spread. Its lower CRPS therefore does not by itself show that it represents uncertainty well.

Table 5: Effects of centering and coarse stochasticity. Bold marks the lowest mean MSE or CRPS.

Dataset	q	Coarse variant	MSE ↓	CRPS ↓	Cov. → .9	Width
Electricity	1	Ours ($\hat{C} + r_s - \bar{r}$)	0.1644 ± 0.0013	0.8667 ± 0.0023	1.0000	10271
		No coarse centering ($\hat{C} + r_s$)	0.3338 ± 0.0054	0.8832 ± 0.0026	1.0000	10271
		Deterministic coarse ($\hat{C}$)	$\mathbf{0.1381 \pm 0.0006}$	$\mathbf{0.2368 \pm 0.0002}$	0.0027	1.3346
	4	Ours ($\hat{C} + r_s - \bar{r}$)	$\mathbf{0.0982 \pm 0.0004}$	0.8576 ± 0.0023	1.0000	10247
		No coarse centering ($\hat{C} + r_s$)	0.2668 ± 0.0040	0.8742 ± 0.0025	1.0000	10247
		Deterministic coarse ($\hat{C}$)	$\mathbf{0.0982 \pm 0.0004}$	$\mathbf{0.1875 \pm 0.0005}$	–	–
Traffic	1	Ours ($\hat{C} + r_s - \bar{r}$)	0.4855 ± 0.0030	1.1141 ± 0.0011	0.9993	0.6662
		No coarse centering ($\hat{C} + r_s$)	0.7180 ± 0.0050	1.1329 ± 0.0014	0.9992	0.6662
		Deterministic coarse ($\hat{C}$)	$\mathbf{0.4122 \pm 0.0013}$	$\mathbf{0.2591 \pm 0.0048}$	0.1004	0.0027
	4	Ours ($\hat{C} + r_s - \bar{r}$)	$\mathbf{0.2144 \pm 0.0007}$	1.0927 ± 0.0015	0.9997	0.6653
		No coarse centering ($\hat{C} + r_s$)	0.4461 ± 0.0068	1.1115 ± 0.0018	0.9997	0.6653
		Deterministic coarse ($\hat{C}$)	$\mathbf{0.2144 \pm 0.0007}$	$\mathbf{0.1983 \pm 0.0016}$	–	–

5.4 Sensitivity to Coarse Diffusion Capacity

Table 6: Sensitivity to the coarse denoiser hidden dimension. Weather and Traffic have 21 and 862 input channels, respectively. Bold marks the lowest mean per dataset.

Dataset	Coarse hidden dimension					
	48	96	192	384	1024	2048
Weather	0.1902 ± 0.0038	0.1865 ± 0.0024	0.1859 ± 0.0007	$\mathbf{0.1832 \pm 0.0010}$	0.1835 ± 0.0018	0.1855 ± 0.0027
Traffic	1.1150 ± 0.0013	1.0844 ± 0.0007	1.0259 ± 0.0045	0.8972 ± 0.0007	0.3814 ± 0.0015	$\mathbf{0.3009 \pm 0.0015}$

The capacity ablation in Table 6 shows that Traffic performance is sensitive to coarse-denoiser capacity. With the detail model fixed, increasing the coarse hidden width from 48 to 2048 reduces Traffic CRPS from 1.1150 to 0.3009. Weather is less sensitive, with its lowest mean CRPS at width 384. These results identify coarse-model capacity as an important factor, but do not establish input dimensionality as the sole cause of the observed sensitivity.

6 Conclusion

We introduced an anchor-preserving coarse-to-fine diffusion framework that retains aggregate variation. Ensemble centering and projected detail preserve the prescribed coarse ensemble mean and sample-wise coherence up to numerical precision. The model achieves the lowest mean fine-resolution CRPS among completed comparisons in 15 of 24 settings, but substantially underperforms on Electricity and Traffic. Controlled comparisons provide limited support for pre-centering in the evaluated settings and illustrate that fine-resolution gains need not extend to aggregate forecasts under matched coarse means. These results support evaluating structural consistency and uncertainty quality separately, with accurate coarse-distribution modeling remaining a key requirement.

AI USE STATEMENT

We used ChatGPT, Codex, and Claude Code to help develop the methodology, formulate mathematical claims, and write mathematical expressions in the proofs. We also used these tools to implement the method, manage experiments, and assist with result interpretation and translation during manuscript preparation.

We did not use generative AI tools to generate synthetic datasets, provide key ideas or arguments for the proofs, propose or refine hypotheses, clean or reformat datasets, or perform qualitative or thematic data analysis. Additionally, we used these tools to search for relevant literature and to draft and edit parts of the manuscript.

The authors reviewed all AI-assisted work.

REPRODUCIBILITY STATEMENT

Section 3 describes the model, learning objectives, and forecast construction. Section 4.1 describes datasets, metrics, and implementation details. Further implementation details can be found in Appendix A. Additionally, an anonymized implementation is included in the supplementary material and will also be released on GitHub.

REFERENCES

George Athanasopoulos, Rob J Hyndman, Nikolaos Kourentzes, and Fotios Petropoulos. Forecasting with temporal hierarchies. *European Journal of Operational Research*, 262(1):60–74, 2017.

Xinyao Fan, Yueying Wu, Chang Xu, Yu-Hao Huang, Weiqing Liu, and Jiang Bian. MG-TSD: Multi-granularity time series diffusion models with guided learning process. In *International Conference on Learning Representations*, volume 2024, pp. 12916–12934, 2024.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Florentin Guth, Simon Coste, Valentin De Bortoli, and Stephane Mallat. Wavelet score-based generative modeling. *Advances in Neural Information Processing Systems*, 35:478–491, 2022.

Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems*, 35:18100–18115, 2022.

Paula Harder, Alex Hernandez-Garcia, Venkatesh Ramesh, Qidong Yang, Prasanna Sattegeri, Daniela Szwarcman, Campbell Watson, and David Rolnick. Hard-constrained deep learning for climate downscaling. *Journal of Machine Learning Research*, 24(365):1–40, 2023.

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

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pp. 8867–8887, 2022.

Christian L Keppenne. An ensemble recentering kalman filter with an application to argo temperature data assimilation into the nasa geos-5 coupled model. *Ocean Modelling*, 77:50–55, 2014.

Chih-Yu Lai, Yu-Chien Ning, and Duane S Boning. RDIT: Residual-based diffusion implicit models for probabilistic time series forecasting. *arXiv preprint arXiv:2509.02341*, 2025.

Qi Li, Zhenyu Zhang, Lei Yao, Zhaoxia Li, Tianyi Zhong, and Yong Zhang. Diffusion-based decoupled deterministic and uncertain framework for probabilistic multivariate time series forecasting. In *International Conference on Learning Representations*, 2025.

Yuxin Li, Wenchao Chen, Xinyue Hu, Bo Chen, Baolin Sun, and Mingyuan Zhou. Transformer-modulated diffusion models for probabilistic multivariate time series forecasting. In *International Conference on Learning Representations*, 2024.

Sai Shankar Narasimhan, Shubhankar Agarwal, Litu Rout, Sanjay Shakkottai, and Sandeep Chinchali. Constrained posterior sampling: Time series generation with hard constraints. *Advances in Neural Information Processing Systems*, 38:107923–107963, 2026.

Anastasios Panagiotelis, Puwasala Gamakumara, George Athanasopoulos, and Rob J Hyndman. Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. *European Journal of Operational Research*, 306(2):693–706, 2023.

Syama Sundar Rangapuram, Lucien D Werner, Konstantinos Benidis, Pedro Mercado, Jan Gasthaus, and Tim Januschowski. End-to-end learning of coherent probabilistic forecasts for hierarchical time series. In *International Conference on Machine Learning*, pp. 8832–8843, 2021.

Syama Sundar Rangapuram, Shubham Kapoor, Rajbir Singh Nirwan, Pedro Mercado, Tim Januschowski, Yuyang Wang, and Michael Bohlke-Schneider. Coherent probabilistic forecasting of temporal hierarchies. In *International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 9362–9376, 2023.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*, pp. 8857–8868. PMLR, 2021.

Alessandro Salatiello, Stefan Birr, and Manuel Kunz. Hierarchical time series forecasting via latent mean encoding. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024.

Mohammad Amin Shabani, Amir H. Abdi, Lili Meng, and Tristan Sylvain. Scaleformer: Iterative multi-scale refining transformers for time series forecasting. In *International Conference on Learning Representations*, 2023.

Lifeng Shen and James Kwok. Non-autoregressive conditional diffusion models for time series prediction. In *International Conference on Machine Learning*, pp. 31016–31029. PMLR, 2023.

Lifeng Shen, Weiyu Chen, and James Kwok. Multi-resolution diffusion models for time series forecasting. In *International Conference on Learning Representations*, pp. 49042–49060, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Andrew Stirn, Harm Wessels, Megan Schertzer, Laura Pereira, Neville Sanjana, and David Knowles. Faithful heteroscedastic regression with neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 5593–5613. PMLR, 2023.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations*, 2023.

Weiwei Ye, Zhuopeng Xu, and Ning Gui. Non-stationary diffusion for probabilistic time series forecasting. In *International Conference on Machine Learning*, volume 267, pp. 72112–72130, 2025.

Lorenzo Zambon, Arianna Agosto, Paolo Giudici, and Giorgio Corani. Properties of the reconciled distributions for gaussian and count forecasts. *International Journal of Forecasting*, 40(4):1438–1448, 2024.

Bohan Zhang, Yanfei Kang, Anastasios Panagiotelis, and Feng Li. Optimal reconciliation with immutable forecasts. *European Journal of Operational Research*, 308(2):650–660, 2023.

A Additional Implementation Details

A.1 Architecture and Training

Architecture. The coarse location predictor uses TimeMixer with hidden dimension 32, feed-forward dimension 64, and two mixing layers. It uses two average-pooling stages with a window size and stride of 2, moving-average decomposition with a window size of 25, channel-independent processing, and dropout 0.1. The coarse and detail denoisers are temporal convolutional networks with hidden widths 48 and 64 and three and four timestep-conditioned residual blocks, respectively. Their dilations are $(1, 2, 4)$ and $(1, 2, 4, 8)$. Both use a kernel size 3, group normalization, SiLU activations, and no dropout.

Training. Each component is trained separately with AdamW with a learning rate 10^{-3}, weight decay 10^{-4}, and gradient-norm clipping at 1. We train for at most 30 epochs, use an early-stopping patience of 5, and retain the checkpoint with the lowest validation loss. Training batch sizes are 64 for ETTh1, ETTh2, ETTm1, and Weather, 16 for Electricity, and 8 for Traffic. Training and validation use coarse-target MAE for the location predictor, noise-prediction MSE for coarse diffusion, and projected-detail MSE plus the leakage penalty for detail diffusion. We set $\lambda_{\text{leak}} = 0.01$; each detail loss term is averaged over its corresponding output elements.

A.2 Residual Preparation and Sampling

All splits use stride-one windows whose targets lie entirely within that split. Validation and test histories may include observations from the preceding split. We evaluate every admissible test window and retain incomplete batches. For Weather, -9999 is treated as missing, and imputation uses training-channel means; missing validation or test targets are not allowed.

Cross-fitting uses five contiguous folds of training-window starts. Each fold fit excludes windows whose full history–target span overlaps the held-out span. Fold predictors use the common validation split for early stopping, not the held-out fold. The final location predictor is fitted on all training windows. Diffusion training uses out-of-fold locations, whereas diffusion validation uses the final predictor. For each coarse lead and channel, m and v are fitted to the N training out-of-fold residuals. We compute the variance with denominator N and floor the resulting standard deviation at 10^{-4}. Both m and v are then frozen.

Both diffusion models use $T = 100$ training steps and a linear β_t schedule from 10^{-4} to 10^{-1}. Each sampler makes 20 denoiser calls at $t = 1 + \mathrm{round}(99(1 - i/19))$, for $i = 0, \ldots, 19$. Code index j corresponds to timestep $t = j + 1$; the clean state $t = 0$ is outside this grid. Starting from $Z_T \sim \mathcal{N}(0, I)$, define $\widehat{\epsilon}_t = \epsilon_\psi(\widehat{C} + Z_t, t; X, \widehat{C})$. For consecutive reverse times $t > \tau > 0$, the coarse update is

$$\begin{aligned} \widehat{Z}_{0|t} &= \frac{Z_t - \sqrt{1 - \bar{\alpha}_t}\,\widehat{\epsilon}_t}{\sqrt{\bar{\alpha}_t}}, \\ Z_\tau &= \sqrt{\bar{\alpha}_\tau}\,\widehat{Z}_{0|t} + \sqrt{1 - \bar{\alpha}_\tau}\,\widehat{\epsilon}_t. \end{aligned} \tag{16}$$

At the final call, $t = 1$, the sampler returns $\widehat{Z}_{0|1}$ without another denoiser call. Each residual is inverse-normalized before centering over all $S = 100$ members for the same history; there is no chunk-wise centering. Detail sampling conditions on the centered coarse scenario and projects the initial state, clean prediction, noise estimate, and every reverse update. Neither sampler adds reverse noise, clips samples, or rescales spread; numerical clamps are used only for stability.

A.3 Evaluation Metrics

Scores, point forecasts, and intervals use the same $S = 100$ samples. At resolution q, we average each sample and target over nonoverlapping q-step blocks aligned to the forecast origin. Quantiles are recomputed from the aggregated samples, rather than by aggregating interval endpoints. All reported aggregation factors divide H. For a scalar target z and sample values $\widehat{z}_s$, the marginal

empirical CRPS is

$$\widehat{\mathrm{CRPS}} = \frac{1}{S}\sum_{s=1}^{S}|\widehat{z}_s - z| - \frac{1}{2S^2}\sum_{s=1}^{S}\sum_{r=1}^{S}|\widehat{z}_s - \widehat{z}_r|. \tag{17}$$

This V-statistic is computed in float64 and averaged with equal weight over all windows, leads, and channels. The point forecast is the sample mean. Central 90% intervals use the 0.05 and 0.95 empirical quantiles with linear interpolation. Coverage includes targets on the interval boundaries, and width is the mean endpoint difference.

CRPS, MAE, and MSE are computed after channel-wise standardization using training-split means and population standard deviations. Standardization precedes temporal aggregation. In Table 5, interval width is reported in original data units after inverse-transforming the samples with the same frozen scaler. Coverage is unit-free. Each metric is computed separately per training seed. Unless stated otherwise, we report the mean and sample standard deviation over seeds 1–3, using variance denominator $3 - 1$. Incomplete seed groups are excluded from complete-group rankings and point-forecast figures.

A.4 Diagnostics of mr-Diff Underdispersion

Across the 72 dataset–horizon–seed settings, our mr-Diff reimplementation has a median central 90% interval coverage of 0.012. We examine the denoising target and future mixup in a targeted experiment with the same five-stage architecture.

Table 7: Diagnostic of our mr-Diff reimplementation on ETTh1 ($H = 96$). All variants use training seed 42, 30 fixed training epochs, 384 test windows, and 50 samples per window. MAE and CRPS are reported in standardized units. Nominal coverage is 0.90.

Target	Future mixup	MAE ↓	CRPS ↓	Coverage
x_0	On	0.558	0.536	0.077
x_0	Off	1.175	0.783	0.512
ϵ	On	0.536	0.454	0.329
ϵ	Off	0.455	0.338	0.737

Replacing clean-target prediction with noise prediction and removing future mixup improves coverage from 0.077 to 0.737 and reduces CRPS from 0.536 to 0.338. Removing mixup alone worsens MAE and CRPS. These results suggest that the two training choices contribute to the observed undercoverage in this setting. However, this single-seed diagnostic neither verifies reproduction of the original paper's point-forecast results nor explains the extreme ETTm1 forecasts. The modified variants are diagnostic controls, not replacements for the mr-Diff baseline in the main comparison.

B Additional Experimental Results

B.1 Point Forecast Results

Figures 2 and 3 report the MAE and MSE of the 100-sample ensemble means. Both metrics are computed after channel-wise standardization fitted on the training split. On the ETT datasets and Weather, MAE follows the CRPS ranking: Ours attains the lowest MAE in 13 of the 16 settings and stays within 1.2% of the best baseline in the remaining three. Its MSE is the lowest on ETTh2 at three of four horizons and otherwise within 2.6% of the best baseline. On Electricity and Traffic, Ours trails TM+G and D^3U by 15–32% in both MAE and MSE but remains ahead of the diffusion baselines NsDiff and mr-Diff at all but one horizon. This point-forecast gap is far smaller than the roughly fivefold CRPS gap on the same datasets, which suggests that the probabilistic deficit there stems mainly from an over-dispersed predictive distribution rather than from the location of the forecast.

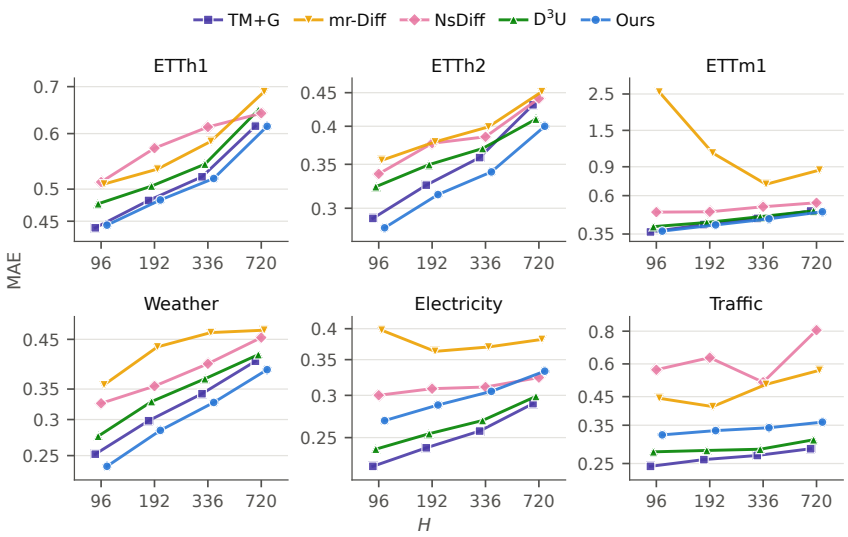

Figure 2: Point forecast MAE (↓) of the 100-sample ensemble mean for the models of Table 2, in train-split standardized units and averaged over seeds 1–3. The vertical axis is logarithmic.

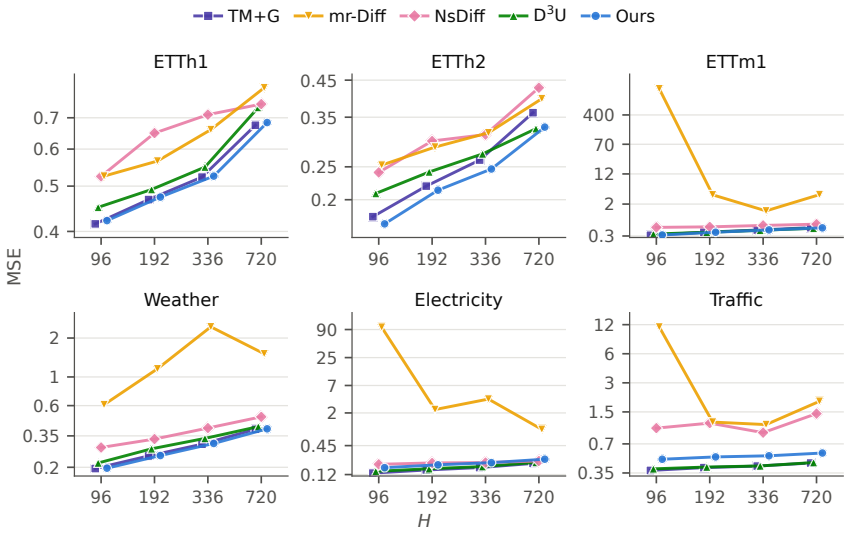

Figure 3: Point forecast MSE (↓), as in Figure 2. The large mr-Diff values come from a few extreme forecasts in individual seeds.

B.2 Visualization of Forecast Results

We visualize ETTh1 forecasts ($L = 336$, $H = 96$, 100 samples) on the HUFL channel. To avoid hand-picking examples, windows are chosen by a fixed rule over all 3,390 test windows. For each window, we average each method's CRPS over HUFL and OT. We then subtract our model's score from the lowest of the four baseline scores. We select windows at the 90th, 50th, and 10th percentiles of this score difference to represent favorable, typical, and unfavorable cases. Because the best baseline is chosen separately for each window, this comparison uses the best baseline separately for each window: Ours outperforms the per-window best baseline on 40.6% of the windows, although it has the lowest mean CRPS over the test set.

In the favorable window (Figure 4), Ours attains the lowest HUFL CRPS (0.377, compared with 0.402 for TM+G). Its coarse anchor underestimates the depth of the recurring drops, but the coarse

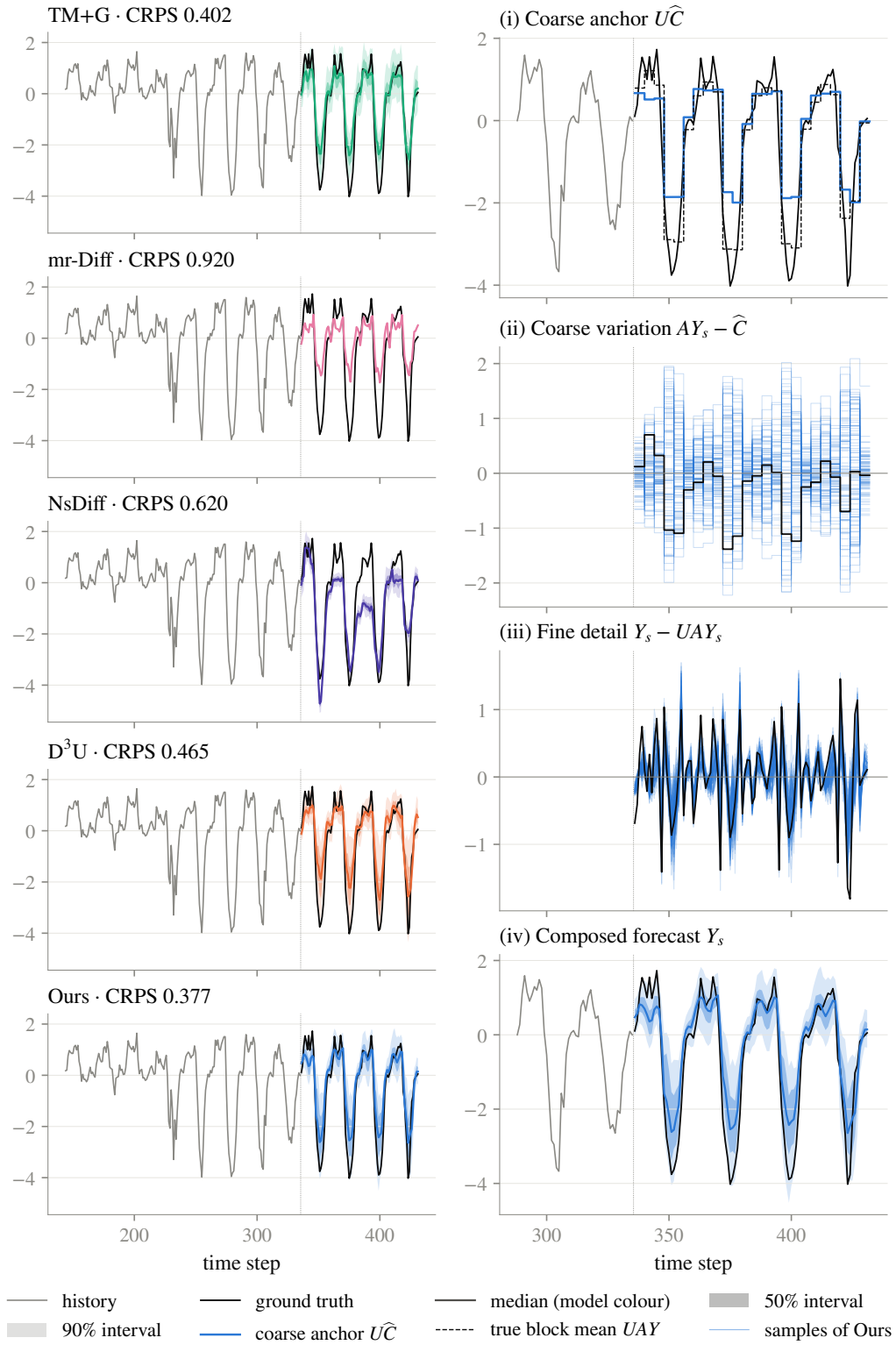

Figure 4: HUFL forecasts for a favorable ETTh1 window. Left: all methods, with the sample median and the central 50% and 90% intervals; gray is the history and black the ground truth. Right: our forecast decomposed into (i) the coarse anchor, (ii) the anchored coarse variation, (iii) the projected fine detail, and (iv) the composed forecast.

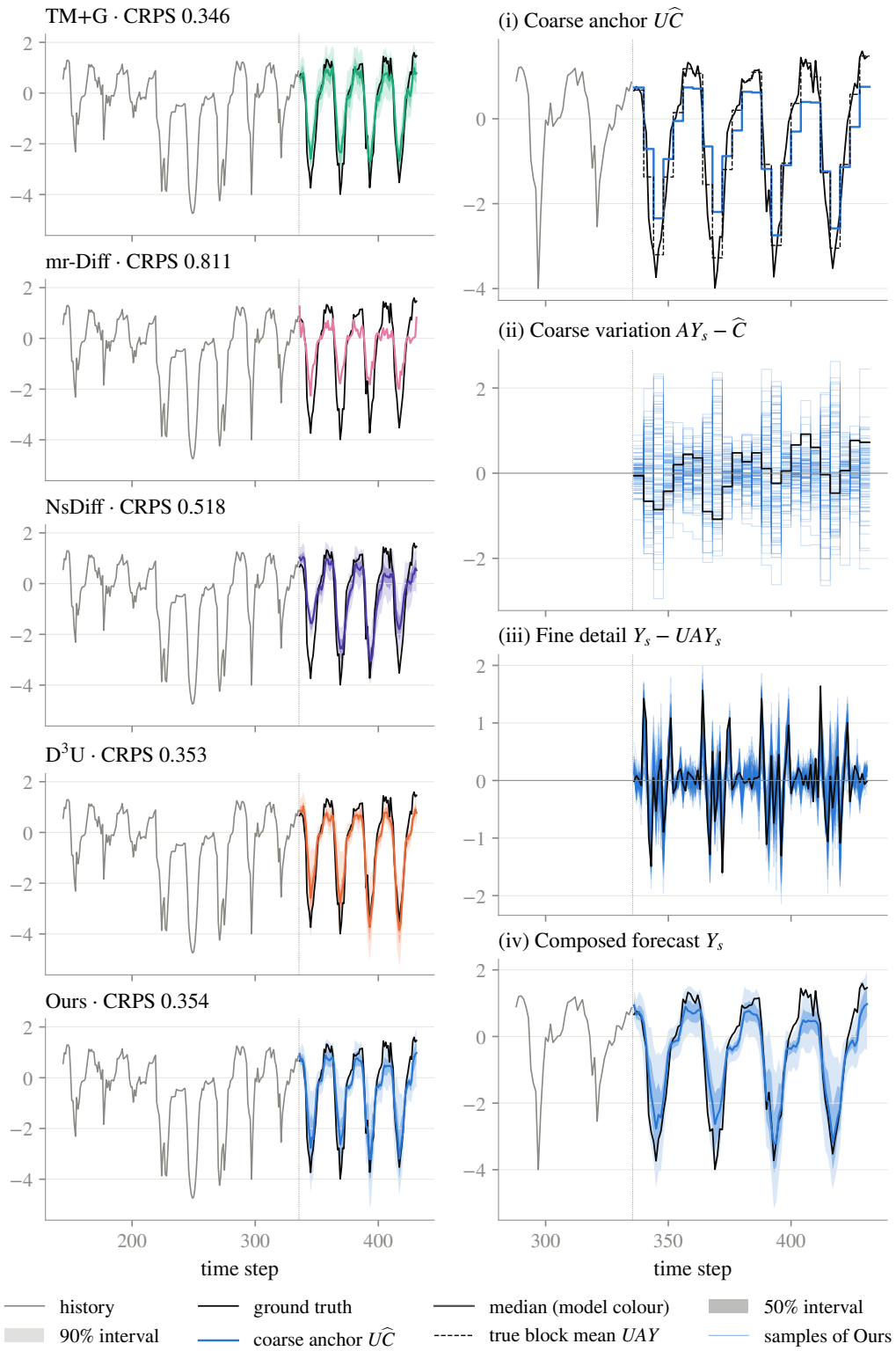

Figure 5: HUFL forecasts for a typical ETTh1 window, drawn as in Figure 4.

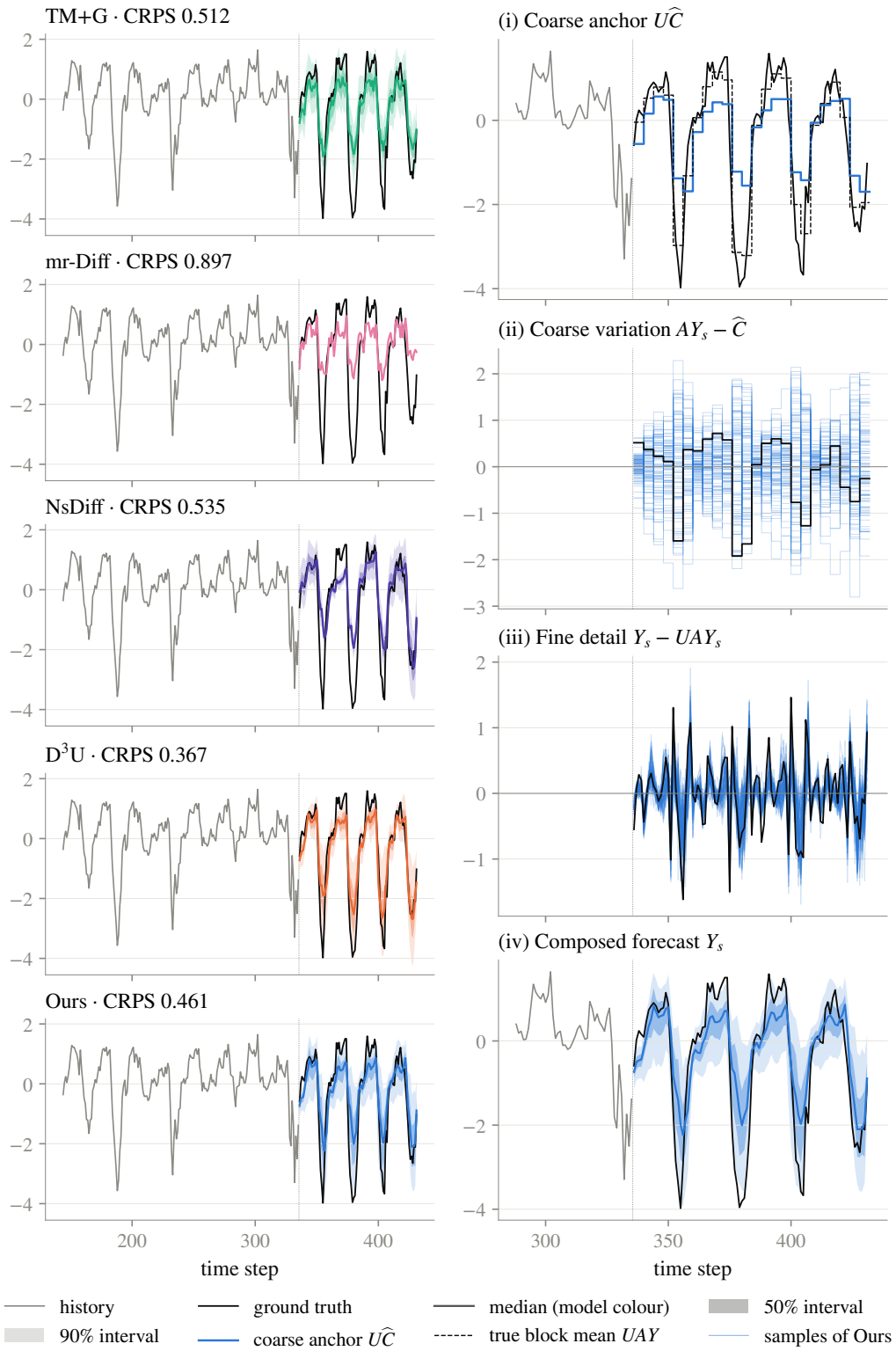

Figure 6: HUFL forecasts for an unfavorable ETTh1 window, drawn as in Figure 4.

variation is widest on those blocks, so the 90% interval contains the ground truth at most forecast steps. In the typical window (Figure 5), TM+G, D^{3}U, and Ours are close (0.346, 0.353, and 0.354). In the unfavorable window (Figure 6), D^{3}U is better than Ours (0.367 and 0.461), because our anchor is too high at the troughs and the intervals miss some of the observed values. In all three windows the projected detail is not negligible on HUFL. Our mr-Diff reimplementation produces almost no spread, so its intervals collapse onto the median and its CRPS (0.81–0.92) is close to the absolute error of its point forecast.

C Proofs of Structural Properties

The algebraic identities below hold for each ensemble produced by the stated constructions. They do not depend on the accuracy of the learned predictor or diffusion models. Sampling assumptions are stated where they are needed. All temporal operators are applied separately to each channel.

C.1 Projection and Coherence

For one channel, let $A \in \mathbb{R}^{K\times H}$ average nonoverlapping blocks of b time steps, where $H = bK$. The repeat operator satisfies $U = bA^{\mathsf{T}}$. Thus,

$$AU = I_K, \qquad UA = \frac{1}{b}UU^{\mathsf{T}}, \qquad (UA)^{\mathsf{T}} = UA, \qquad (UA)^2 = UA. \tag{18}$$

These identities also hold channelwise in the matrix notation of Section 3.1. Define $\Pi = I_H - UA$. Then

$$\Pi^2 = \Pi, \qquad \Pi^{\mathsf{T}} = \Pi, \qquad A\Pi = A - AUA = 0. \tag{19}$$

For any $d \in \ker(A)$, we have $\Pi d = d$. Together with $A\Pi = 0$, this gives $\mathrm{range}(\Pi) = \ker(A)$. Thus Π is the orthogonal projection onto the null space $\ker(A)$. Signals in this space have zero average in every block.

For $C^\star = AY$ and $D_0 = \Pi Y$,

$$AD_0 = 0, \qquad UC^\star + D_0 = UAY + (I_H - UA)Y = Y. \tag{20}$$

This decomposition is unique. Applying A to any decomposition $Y = UC + D$ with $AD = 0$ gives $C = AY$. It then follows that $D = Y - UAY = \Pi Y$.

The detail forward process also stays in $\ker(A)$. Applying A to Eq. (11) gives

$$AD_t = \sqrt{\bar\alpha_t}\, AD_0 + \sqrt{1-\bar\alpha_t}\, A\Pi\xi = 0, \tag{21}$$

since $AD_0 = 0$ and $A\Pi = 0$.

The same property holds for the detail sampler. Its initial state $D_T = \Pi\xi_T$ lies in $\ker(A)$. At each reverse step, Eq. (14) uses projected quantities:

$$\widehat{D}_{0|t} = \Pi h_\theta(\cdot), \qquad \widehat{\xi}_t = \Pi(\cdot), \qquad D_\tau = \Pi(\cdot).$$

The predicted clean detail, noise estimate, and next state therefore all lie in $\ker(A)$. Induction over the reverse steps gives $AD_s = 0$ for each final detail sample. For $Y_s = UC_s + D_s$,

$$AY_s = AUC_s + AD_s = C_s. \tag{22}$$

Thus projection ensures exact sample-wise coherence, up to floating-point roundoff. This holds regardless of the denoiser output before projection.

For standard Gaussian noise $\xi_{:,d} \sim \mathcal{N}(0, I_H)$ in channel d,

$$\mathrm{Cov}((\Pi\xi)_{:,d}) = \Pi\Pi^{\mathsf{T}} = \Pi.$$

C.2 Ensemble Centering

Let $\bar r = S^{-1}\sum_{s=1}^S r_s$ and $C_s = \widehat{C} + r_s - \bar r$. Direct averaging gives

$$\begin{aligned} \frac{1}{S}\sum_{s=1}^S C_s &= \widehat{C} + \frac{1}{S}\sum_{s=1}^S r_s - \bar r = \widehat{C}, \\ C_s - C_{s'} &= r_s - r_{s'}. \end{aligned} \tag{23}$$

For the covariance calculation, let $\rho_s = \mathrm{vec}(r_s)$ and $c_s = \mathrm{vec}(C_s)$. Write $\widehat{c} = \mathrm{vec}(\widehat{C})$ and $\bar{\rho} = S^{-1}\sum_{s=1}^{S}\rho_s$. For $S > 1$,

$$\frac{1}{S-1}\sum_{s=1}^{S}(c_s - \widehat{c})(c_s - \widehat{c})^{\mathsf{T}} = \frac{1}{S-1}\sum_{s=1}^{S}(\rho_s - \bar{\rho})(\rho_s - \bar{\rho})^{\mathsf{T}}. \tag{24}$$

Thus centering shifts the coarse ensemble but preserves its pairwise differences and empirical covariance.

Centering does not generally preserve the joint sampling distribution. Suppose that $\rho_1, \ldots, \rho_S$ are independent conditional on X. Assume also that $\mathrm{Cov}(\rho_s \mid X) = \Sigma$ for every s. Then

$$\begin{aligned}\mathrm{Cov}(\rho_s - \bar{\rho} \mid X) &= (1 - S^{-1})\Sigma,\\ \mathrm{Cov}(\rho_s - \bar{\rho}, \rho_{s'} - \bar{\rho} \mid X) &= -S^{-1}\Sigma, \qquad s \neq s'.\end{aligned}$$

The centered members are therefore generally dependent. These identities alone do not imply calibration.

Combining Eq. (22) with the mean identity in Eq. (23) gives

$$\begin{aligned}A\bar{Y} &= \frac{1}{S}\sum_{s=1}^{S} AY_s = \widehat{C},\\ \bar{Y} &= U\widehat{C} + \bar{D}, \qquad \bar{D} = \frac{1}{S}\sum_{s=1}^{S} D_s, \qquad A\bar{D} = 0.\end{aligned} \tag{25}$$

Thus the fine-resolution mean $\bar{Y}$ need not equal $U\widehat{C}$, because $\bar{D}$ can be nonzero. Only its aligned block averages are fixed.

C.3 Aligned Temporal Aggregation

Let $q = \ell b$, where ℓ is a positive integer and q divides H. Let A_q average aligned, nonoverlapping blocks of q fine time steps. Define B_ℓ to average ℓ consecutive coarse time steps. Each q-block contains exactly ℓ complete b-blocks. Direct summation therefore gives

$$A_q U = B_\ell, \qquad A_q = B_\ell A. \tag{26}$$

If $AD = 0$, this factorization gives $A_q D = B_\ell A D = 0$. Applying these identities to the generated samples yields

$$A_q Y_s = B_\ell C_s, \qquad A_q \bar{Y} = B_\ell \widehat{C}. \tag{27}$$

With $b = 4$, this covers $q = 4, 12, 24$ (with $\ell = 1, 3, 6$), provided that q divides H.

If q is not a multiple of b, an aggregation window includes partial b-blocks. The same can happen if its boundaries are shifted off the b-block grid. In these cases, the details need not cancel.

For a deterministic coarse variant, suppose $C_s = \widehat{C}$ for every member. Then every aligned aggregate satisfies $A_q Y_s = B_\ell \widehat{C}$. The aggregate ensemble has zero spread, even though the fine-resolution details may vary within each block.

C.4 Post-hoc Mean Correction

Let $\{F_s\}_{s=1}^{S}$ be any completed fine-resolution ensemble, and let $\bar{F} = S^{-1}\sum_{s=1}^{S} F_s$. Apply the same correction to each member:

$$\Delta = \widehat{C} - A\bar{F}, \qquad F'_s = F_s + U\Delta. \tag{28}$$

Using $AU = I_K$ and linearity,

$$\begin{aligned}A\bar{F}' &= A\bar{F} + AU\Delta = \widehat{C},\\ F'_s - \bar{F}' &= F_s - \bar{F},\\ F'_s - F'_{s'} &= F_s - F_{s'}.\end{aligned} \tag{29}$$

The correction preserves pairwise differences and empirical covariance after vectorization. At any fixed linear aggregation, all members receive the same shift. The widths of intervals based on empirical quantiles are therefore unchanged. The ensemble's location relative to the observed target can still change.

The correction also leaves the within-block details unchanged. Using $I_H = UA + \Pi$, we can write

$$F'_s = U\left(\widehat{C} + AF_s - A\bar{F}\right) + \Pi F_s. \tag{30}$$

Now consider an uncentered ensemble

$$\widetilde{F}_s = U(\widehat{C} + r_s) + D_s, \qquad AD_s = 0.$$

Its coarse mean is $A\overline{\widetilde{F}} = \widehat{C} + \bar{r}$, so the correction is $-U\bar{r}$. Hence,

$$\widetilde{F}_s + U\left(\widehat{C} - A\overline{\widetilde{F}}\right) = U(\widehat{C} + r_s - \bar{r}) + D_s. \tag{31}$$

For fixed detail samples, this is exactly the ensemble obtained by centering the coarse residuals before adding those details.

For aligned $q = \ell b$, Eq. (26) also gives $A_q\bar{F}' = B_\ell\widehat{C}$. Both procedures therefore give the same guarantee for the coarse ensemble mean.

This mean guarantee does not imply that the two sampling procedures are equivalent. Post-hoc correction keeps the existing details ΠF_s. The proposed sampler generates details conditional on the centered C_s. The two procedures need not produce the same distribution.