Video Generation Empowered Long-Term Radio Map Prediction in UAV-Assisted Communication

Anonymous Authors[1]

Abstract

Radio maps provide rich information about the radio landscape that is useful to a myriad of wireless communication and sensing applications. In contrast to the scenario with fixed-position transmitters, the transmitter in the unmanned aerial vehicle (UAV)-assisted communication scenario can move along a predefined trajectory, resulting in a continuously evolving *long-term radio map*. To address this long-term radio map prediction problem, existing deep learning-based methods that follow the image-to-image translation approach suffer from significant complexity and limited accuracy due to their independent frame-by-frame construction method, which neglects the temporal correlation of radio characteristics. In this work, we formulate the long-term radio map prediction problem in the UAV-assisted communication system as a *conditioned video generation* task. In particular, we propose a *lightweight* model that encodes the static environment only once, conditions radio map prediction on both the environment and the UAV trajectory, and refines the generated sequence in the temporal domain by exploiting double-sided dependencies. Experiments on a large-scale simulated dataset demonstrate that the proposed method consistently improves prediction accuracy and temporal consistency while reducing computational redundancy compared with both static baselines and sequential baselines. The source code is available at GitHub repository.

1. Introduction

A radio map characterizes the spatial distribution of radio-frequency quantities, such as channel path loss, in the wireless environments (Zhang et al., 2024; Bi et al., 2019; Romero & Kim, 2022; Zeng et al., 2024). By capturing the radio propagation landscape, radio maps enable numerous wireless applications, including spectrum management in cognitive radio networks (Bi et al., 2019; Romero & Kim, 2022), localization (Yapar et al., 2023), and unmanned aerial vehicle (UAV) path planning (Zhang et al., 2019; Zhang & Zhang, 2021).

Most existing deep learning (DL)-based radio map construction methods are developed for static configurations with fixed-position transmitters. In particular, RadioUNet (Levie et al., 2021) formulates radio map estimation as an *image-to-image translation* problem and demonstrates that the U-Net architecture (Ronneberger et al., 2015) can accurately predict radio maps from environment maps. Building on this idea, subsequent studies have explored more advanced generative models, including GAN-based method (Zhang et al., 2023) and diffusion model (Wang et al., 2025), to further improve prediction accuracy. However, these approaches are restrictive in emerging UAV-assisted wireless systems, where the transmitter is mounted on a moving aerial platform that follows a predefined trajectory (Zeng et al., 2019; Wu et al., 2018). In such scenarios, the radio field over the service area evolves continuously with the UAV position and exhibits strong temporal correlations. Consequently, the desired output is a sequence of radio maps in *video form*, where each frame corresponds to the radio map associated with a particular UAV position. We refer to such a radio map sequence as a *long-term radio map* in this work.

The long-term radio maps are valuable because many downstream tasks depend not only on instantaneous coverage, but also on its temporal evolution. For example, terrestrial base stations may use the radio map sequence to adapt beam steering, cooperative transmission, or handover as the UAV moves, while a network controller may exploit it to anticipate coverage holes, schedule users proactively, or evaluate candidate UAV trajectories prior to deployment (Zhang & Zhang, 2021; Hoffmann & Kryszkiewicz, 2023).

A straightforward way to construct such a long-term radio map is to apply the static estimators in (Levie et al., 2021; Zhang et al., 2023; Wang et al., 2025) independently at every waypoint along the trajectory. However, this frame-by-frame strategy is suboptimal because it repeatedly pro-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

cesses the same static environment and fails to exploit the strong correlation between nearby UAV positions, where the radio map should typically vary smoothly. As a result, it incurs unnecessary computational cost and may produce temporally inconsistent predictions, which can in turn lead to unstable downstream decisions.

In this paper, we study the long-term radio map prediction problem for UAV-assisted communication, where a UAV-mounted transmitter moves along a predefined trajectory over a static environment. Given an environment map and a UAV trajectory, we propose to solve the problem via the paradigm of *conditioned video generation*. In particular, we propose a *lightweight* network that encodes the static environment once, conditions radio map prediction on the UAV trajectory and the static environment, and refines the generated video in the temporal domain by exploiting double-sided dependencies. Experimental results show that the proposed model improves prediction accuracy, produces temporally consistent radio map transitions, and achieves better amortized inference efficiency than both frame-by-frame static baselines and sequential baselines.

2. Related Work

Temporally-correlated radio map estimation. Recent work has investigated temporally-correlated radio map estimation induced by moving obstacles, in which a ConvLSTM-based autoregressive model is trained to forecast future radio maps from a sequence of historically observed radio maps (Cheng et al., 2026). However, the autoregressive design suffers from two major limitations. First, since each radio map in the sequence is generated sequentially, the inference latency scales linearly with the prediction horizon. Second, the model exploits only *single-sided* temporal dependencies from the historical observations to the current prediction, which restricts its representational capacity. In contrast, our model generates the whole radio map sequence simultaneously in a single forward pass. This design admits parallel processing across frames, thereby reducing the amortized latency. Moreover, it enables the exploitation of *double-sided* temporal dependencies, in which each frame is informed by both past and future frames, thereby enhancing representational capacity and improving overall prediction accuracy (Zhu et al., 2023).

Conditioned video generation. Our problem can be naturally viewed as conditioned video generation, where the condition consists of the static environment and the UAV trajectory. Recent advances in video generation, especially diffusion-based models, have shown strong capability in producing temporally coherent sequences (Ho et al., 2022). However, state-of-the-art video generation models typically incur high computational cost due to large model sizes or iterative sampling procedures (Blattmann et al., 2023), making them less suitable for deployment in wireless systems with strict latency and resource constraints. Instead, we advocate for a lightweight model design that leverages problem-specific inductive biases to achieve a better balance between accuracy and computational efficiency.

3. Problem Formulation

Trajectory-induced long-term radio map. We consider a UAV-assisted communication scenario where a transmitter is mounted on a UAV flying along a predefined trajectory $\mathbf{Q} = [\mathbf{q}_1, \cdots, \mathbf{q}_T] \in \mathbb{R}^{2\times T}$ for safety consideration, where $\mathbf{q}_t \in \mathbb{R}^{2\times 1}$ represents the horizontal location of the UAV at time slot t, $\forall t\{1, \cdots, T\}$.[1] Consider a region of interest $\mathcal{R} \subset \mathbb{R}^2$ discretized into $H \times W$ grids. The static propagation environment is represented by a bird's-eye-view building-height map $\mathbf{E} \in \mathbb{R}^{H\times W}$, where the (i, j)-th entry $[\mathbf{E}]_{i,j}$ denotes the building height if grid (i, j) is occupied by a building and zero otherwise. At each time slot t, the radio map is determined by both the environment $\mathbf{E}$ and the UAV location $\mathbf{q}_t$, and is denoted by $\mathbf{P}(\mathbf{E}, \mathbf{q}_t) \in \mathbb{R}^{H\times W}$, whose entries represent the path loss from the transmitter to the associated grid center. Further details on the path loss definition and the environment-aware radio map formulation can be found in (Zhu et al., 2026a;b).

Stacking the radio maps across all time slots yields a long-term radio map $\mathbf{V}(\mathbf{E}, \mathbf{Q}) \in \mathbb{R}^{T\times H\times W}$, where the t-th frame is given by $[\mathbf{V}(\mathbf{E}, \mathbf{Q})]_{t,:,:} = \mathbf{P}(\mathbf{E}, \mathbf{q}_t)$. Since both the environment $\mathbf{E}$ and the trajectory $\mathbf{Q}$ are known a priori, the entire sequence $\mathbf{V}(\mathbf{E}, \mathbf{Q})$ can be predicted before the UAV is deployed.

Learning task. To exploit the spatiotemporal correlation between consecutive radio map frames, we formulate the problem as a conditioned video generation task. This formulation departs from conventional radio map estimators, which predict $\mathbf{P}(\mathbf{E}, \mathbf{q}_t)$ independently for each transmitter location and therefore ignores temporal coherence (Levie et al., 2021; Zhang et al., 2023; Wang et al., 2025). Formally, given the static environment $\mathbf{E}$ and a UAV trajectory $\mathbf{Q}$, the goal is to learn a mapping f_θ that generates the entire sequence of radio maps:

$$f_\theta : (\mathbf{E}, \mathbf{Q}) \longmapsto \widehat{\mathbf{V}}(\mathbf{E}, \mathbf{Q}) \in \mathbb{R}^{T\times H\times W}, \tag{1}$$

where $\widehat{\mathbf{V}}(\mathbf{E}, \mathbf{Q})$ denotes the predicted long-term radio map. Given a training set $\mathcal{D} = \{(\mathbf{E}^{(n)}, \mathbf{Q}^{(n)}, \mathbf{V}^{(n)})\}_{n=1}^{N}$, where $\mathbf{V}^{(n)} \equiv \mathbf{V}(\mathbf{E}^{(n)}, \mathbf{Q}^{(n)})$, the model parameters θ are learned

[1] We assume a fixed UAV altitude higher than the maximum building height, so that the transmitter location is fully specified by its horizontal coordinates (Wu et al., 2018).

by minimizing

$$\mathcal{L}_\theta = \frac{1}{N}\sum_{n=1}^{N}\left[\mathcal{L}_{\rm rec}\big(\widehat{\mathbf{V}}^{(n)},\mathbf{V}^{(n)}\big)+\lambda\,\mathcal{L}_{\rm tc}\big(\widehat{\mathbf{V}}^{(n)},\mathbf{V}^{(n)}\big)\right]. \tag{2}$$

where $\widehat{\mathbf{V}}^{(n)} = f_\theta(\mathbf{E}^{(n)},\mathbf{Q}^{(n)})$ and $\lambda \geq 0$ balances per-frame reconstruction accuracy and temporal consistency. The reconstruction term $\mathcal{L}_{\rm rec}(\cdot,\cdot)$ is defined as the mean-squared error over all frames and grids:

$$\mathcal{L}_{\rm rec}(\widehat{\mathbf{V}},\mathbf{V}) = \frac{1}{THW}\sum_{t=1}^{T}\left\|\widehat{\mathbf{V}}_t-\mathbf{V}_t\right\|_F^2, \tag{3}$$

where $\mathbf{V}_t = [\mathbf{V}(\mathbf{E},\mathbf{Q})]_{t,:,:}$ and $\widehat{\mathbf{V}}_t = [\widehat{\mathbf{V}}(\mathbf{E},\mathbf{Q})]_{t,:,:}$ denote the t-th frame of ground-truth and predicted radio map sequence, respectively. To encourage temporal coherence, $\mathcal{L}_{\rm tc}(\cdot,\cdot)$ matches the first-order temporal differences between consecutive radio map frames. Let $\Delta\mathbf{V}_t = \mathbf{V}_{t+1}-\mathbf{V}_t$ and $\Delta\widehat{\mathbf{V}}_t = \widehat{\mathbf{V}}_{t+1}-\widehat{\mathbf{V}}_t$ denote the frame-to-frame variations of the ground-truth and predicted long-term radio maps at time slot $t \in \{1,\cdots,T-1\}$, respectively. Then, $\mathcal{L}_{\rm tc}(\cdot,\cdot)$ can be expressed as

$$\mathcal{L}_{\rm tc}(\widehat{\mathbf{V}},\mathbf{V}) = \frac{1}{(T-1)HW}\sum_{t=1}^{T-1}\left\|\Delta\widehat{\mathbf{V}}_t-\Delta\mathbf{V}_t\right\|_F^2. \tag{4}$$

4. Methodology

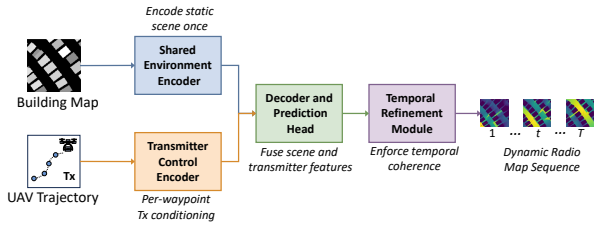

Figure 1. High-level overview of the proposed model for long-term radio map construction. A detailed module-level architecture is provided in Appendix A.

Architecture overview. We instantiate the mapping f_θ as a spatiotemporal encoder-decoder network. The key idea is to separate *time-invariant* information, i.e., the static environment encoded by the building map, from *time-varying* information, i.e., the transmitter location along the UAV trajectory. This design allows the static environment to be encoded only once and reused across all time slots, thereby reducing redundant computation. As shown in Figure 1, the proposed model consists of four main modules: i) a *shared environment encoder*, which extracts a multi-scale feature pyramid from the static building map, ii) a *transmitter control encoder*, which converts each waypoint into transmitter-dependent conditioning features, iii) a *decoder and prediction head*, which fuses these features to reconstruct the long-term radio map, and iv) a *temporal refinement module*, which captures inter-frame dependencies in a compact latent space by considering double-sided temporal correlations. The following paragraphs describe these modules in detail

Static scene encoder. The static scene encoder aims to extract a reusable multi-scale representation of the environment from the building map $\mathbf{E}$. To explicitly model the sharp geometric transitions, we augment $\mathbf{E}$ with a fixed edge map computed using Sobel operators:

$$\mathbf{G}(\mathbf{E}) = \sqrt{(\mathbf{S}_x*\mathbf{E})^2+(\mathbf{S}_y*\mathbf{E})^2+\epsilon}, \tag{5}$$

where $*$ denotes convolution and $\epsilon > 0$ is a small constant for numerical stability. The resulting two-channel input $\mathrm{Concat}_c(\mathbf{E},\mathbf{G}(\mathbf{E}))$ is then passed through a residual convolutional encoder with progressive downsampling, where $\mathrm{Concat}_c(\cdot,\cdot)$ denotes channel-wise concatenation. Specifically, an initial residual stem produces full-resolution features, followed by a sequence of downsampling blocks that generate increasingly coarse representations with larger receptive fields. This yields a scene feature pyramid

$$\left\{\mathbf{S}^{(0)},\mathbf{S}^{(1)},\ldots,\mathbf{S}^{(L)}\right\} = \Phi_{\rm scene}\left(\mathrm{Concat}_c(\mathbf{E},\mathbf{G}(\mathbf{E}))\right), \tag{6}$$

where $\mathbf{S}^{(\ell)}$ denotes the scene feature map at scale ℓ. Because $\mathbf{E}$ is static over time, this scene pyramid is computed once and then broadcast across all T time slots to provide shared spatial context for the subsequent decoding process.

Transmitter-centric control representation. This module converts each UAV waypoint into a dense transmitter-centric representation that provides frame-specific conditions. For each waypoint $\mathbf{q}_t$, we construct a two-channel control map $\mathbf{C}_t \in \mathbb{R}^{2\times H\times W}$ that encodes the transmitter location in image form. The first channel is a Gaussian heatmap $\mathbf{C}_t^{\rm heat} \in \mathbb{R}^{H\times W}$ centered at the transmitter:

$$[\mathbf{C}_t^{\rm heat}]_{i,j} = \exp\left(-\frac{\|\mathbf{g}_{ij}-\mathbf{q}_t\|_2^2}{2\sigma^2}\right), \tag{7}$$

where $\mathbf{g}_{ij}$ denotes the centroid of grid (i,j) and σ controls the spatial spread. The second channel is a normalized log-distance map $\mathbf{C}_t^{\rm dist} \in \mathbb{R}^{H\times W}$:

$$[\mathbf{C}_t^{\rm dist}]_{i,j} = \frac{\log\big(1+\alpha\|\mathbf{g}_{ij}-\mathbf{q}_t\|_2\big)}{\log(1+\alpha d_{\max})}, \tag{8}$$

where α is a scaling constant and $d_{\max}$ is the maximum distance in the normalized grid.

Each control map is then processed independently by a lightweight multi-scale encoder:

$$\left\{\mathbf{U}_t^{(0)},\mathbf{U}_t^{(1)},\ldots,\mathbf{U}_t^{(L)}\right\} = \Phi_{\rm ctrl}(\mathbf{C}_t), \tag{9}$$

where $\mathbf{U}_t^{(\ell)}$ denotes the control feature map at scale ℓ. The encoder is designed such that $\mathbf{U}_t^{(\ell)}$ is spatially aligned with the scene feature map $\mathbf{S}^{(\ell)}$, enabling feature fusion at corresponding resolutions in the decoder. To further obtain a compact summary of the transmitter state at time slot t, we apply global average pooling (GAP) to the multi-scale control features and concatenate the resulting vectors:

$$\mathbf{z}_t = \mathrm{Concat}_{\ell=0}^{L}\big(\mathrm{GAP}(\mathbf{U}_t^{(\ell)})\big), \tag{10}$$

where $\mathrm{Concat}_{\ell=0}^{L}(\cdot)$ denotes the ordered concatenation over the scale index ℓ. The descriptor $\mathbf{z}_t$ is later used to modulate decoder features, while the control pyramid $\{\mathbf{U}_t^{(\ell)}\}_{\ell=0}^{L}$ provides transmitter-dependent condition at each resolution.

Decoder and prediction head. The decoder reconstructs the long-term radio map in a coarse-to-fine manner by combining three complementary information streams: the shared scene pyramid $\{\mathbf{S}^{(\ell)}\}_{\ell=0}^{L}$, the control pyramid $\{\mathbf{U}_t^{(\ell)}\}_{\ell=0}^{L}$, and the compact control descriptor $\mathbf{z}_t, \forall t \in \{1 \cdots, T\}$. For each time slot t, the coarsest scene and control features are first concatenated and fused by a bottleneck residual block:

$$\mathbf{H}_t^{(L)} = \Psi_{\mathrm{bot}}\big(\mathrm{Concat}_c(\mathbf{S}^{(L)}, \mathbf{U}_t^{(L)})\big), \tag{11}$$

The decoder then applies a sequence of upsampling-fusion operators. At scale ℓ, the current latent feature is bilinearly upsampled and projected, modulated by feature-wise affine conditioning driven by $\mathbf{z}_t$ (Perez et al., 2018), and concatenated with the aligned scene and control skip features. A residual fusion block is then used to aggregate these signals:

$$\mathbf{H}_t^{(\ell)} = \Psi_{\mathrm{up}}^{(\ell)}\big(\mathbf{H}_t^{(\ell+1)}, \mathbf{S}^{(\ell)}, \mathbf{U}_t^{(\ell)}, \mathbf{z}_t\big), \quad \ell = L-1, \ldots, 0. \tag{12}$$

In our implementation, the temporal refiner is inserted after the decoder reaches the intermediate scale $\ell = 2$, as described in the next paragraph. The remaining decoder stages subsequently operate on the refined latent representation $\{\widetilde{\mathbf{H}}_t^{(2)}\}_{t=1}^{T}$ to recover full-resolution features. Finally, a lightweight residual prediction head, implemented as a residual block followed by a 1×1 convolution, produces a single-channel correction map $\mathbf{R}_t$. To explicitly capture the coarse distance-dependent attenuation trend, we introduce a learnable radial prior from the normalized log-distance map $\mathbf{C}_t^{\mathrm{dist}}$:

$$\mathbf{M}_t = a\mathbf{C}_t^{\mathrm{dist}} + b\mathbf{1}, \tag{13}$$

where $\mathbf{1} \in \mathbb{R}^{H \times W}$ is an all-ones matrix and $a, b \in \mathbb{R}$ are learned scalar parameters. The final prediction is then given by

$$\widehat{\mathbf{V}}_t = \mathbf{R}_t + \mathbf{M}_t. \tag{14}$$

Temporal refinement in latent space. The temporal refinement module aims to capture inter-frame dependency after the static scene features and transmitter-conditioned features have been fused. Since the UAV trajectory is predefined, we therefore adopt a non-causal refinement strategy that can leverage both past and future context for each frame. In our implementation, temporal refinement is performed at the intermediate decoder scale $\ell = 2$, i.e., at spatial resolution $H/4 \times W/4$. This choice keeps temporal modeling efficient while preserving the key structure needed for accurate reconstruction.

Let $\mathbf{H}_t^{(2)} \in \mathbb{R}^{C_2 \times H/4 \times W/4}$ denote the partially decoded latent feature at time slot t after the first two upsampling stages. Stacking $\{\mathbf{H}_t^{(2)}\}_{t=1}^{T}$ over time yields a sequence tensor $\mathbf{H}_{1:T}^{(2)}$ of dimension $C_2 \times T \times H/4 \times W/4$. We refine this tensor using a lightweight spatiotemporal block. Each block first performs spatial mixing independently at each time step using a depthwise convolution with kernel size $(1, 3, 3)$, followed by a pointwise $1 \times 1 \times 1$ projection. Temporal interaction is then introduced by a symmetric temporal operator

$$\mathcal{T}(\mathbf{X}) = \frac{1}{2}\left(h(\mathbf{X}) + \mathrm{Rev}\big(h(\mathrm{Rev}(\mathbf{X}))\big)\right), \tag{15}$$

where $h(\cdot)$ denotes a depthwise temporal convolution with kernel size k_t followed by pointwise channel mixing, and $\mathrm{Rev}(\cdot)$ reverses the feature sequence along the time axis. After group normalization and GELU activation, the refined feature is added back to the input through a residual connection. Repeating this block a small number of times yields the temporally refined latent sequence

$$\{\widetilde{\mathbf{H}}_t^{(2)}\}_{t=1}^{T} = \Phi_{\mathrm{temp}}\big(\mathbf{H}_{1:T}^{(2)}\big). \tag{16}$$

The refined latent sequence is then passed to the remaining decoder stages to reconstruct the full-resolution radio maps.

5. Experiments

5.1. Baselines

We compare the proposed method with three representative baselines that cover recurrent, volume-based, and per-frame prediction paradigms. All baselines are trained from scratch on the same dataset, using identical data splits, optimizer, and loss function as the proposed model. Details of the dataset construction and splits are provided in Appendix B.

1. **ConvLSTM:** We implement an autoregressive encoder-decoder based on ConvLSTM units as in (Cheng et al., 2026), which predicts the radio map sequence sequentially from the static environment map $\mathbf{E}$ and the per-frame transmitter location.
2. **3D U-Net:** We adopt a naive 3D U-Net that treats the task as direct spatiotemporal volume prediction. It takes $\mathbf{E}$ together with a trajectory-conditioned transmitter tensor as input and outputs the full long-term radio map in a single forward pass.

3. **RadioUNet:** We adapt RadioUNet (Levie et al., 2021) to the dynamic setting by applying it independently at each UAV position and stacking the predictions to form a radio map sequence.

5.2. Metrics

We evaluate the proposed model and the baselines using three groups of metrics: i) *radio map reconstruction fidelity*, including root mean square error (RMSE) and structural similarity index measure (SSIM) (Wang et al., 2004), ii) *temporal-evolution consistency*, measured by delta-map RMSE (DM-RMSE), and iii) *communication-oriented coverage utility*, measured by coverage intersection over union (CIoU). All metrics are computed on denormalized radio maps with path loss values in dB scale.

Radio map reconstruction fidelity. RMSE and SSIM evaluate the per-frame quality of the generated radio map sequence, which quantify the task from a radio map sequence reconstruction perspective. Specifically, RMSE measures the average pixel-wise reconstruction error over space and time, which is reported as the square-root of (3). SSIM is computed frame-wise using the standard window-based implementation (Wang et al., 2004) and then averaged over the trajectory.

Temporal-evolution consistency. In long-term radio map construction, accurate frame-wise prediction is not sufficient. The generated raio maps should also evolve smoothly and consistently as the transmitter moves along its trajectory. Therefore, we report DM-RMSE, the square-root counterpart of (4), to evaluate whether the model preserves the temporal difference patterns between consecutive radio maps.

Communication-oriented coverage utility. Beyond pixel-level reconstruction, radio maps are often used to support coverage analysis and communication-aware decision making. We therefore report CIoU, which measures whether the predicted map correctly identifies the serviceable coverage region under a given path loss threshold. For a threshold $\gamma\, dB$, we define $\mathcal{C}_t^\gamma = \{(i,j) \in \Omega \mid [\mathbf{V}_t]_{i,j} \le \gamma\}$ and $\hat{\mathcal{C}}_t^\gamma = \left\{(i,j) \in \Omega \mid [\widehat{\mathbf{V}}_t]_{i,j} \le \gamma\right\}$ as the ground-truth and predicted coverage regions at time slot t, respectively, where $\Omega = \{1,\dots,H\} \times \{1,\dots,W\}$ denotes the set of grids. The CIoU at threshold γ is then computed as

$$\mathrm{CIoU@}\gamma = \frac{1}{T}\sum_{t=1}^{T} \frac{\left|\mathcal{C}_t^\gamma \cap \hat{\mathcal{C}}_t^\gamma\right|}{\left|\mathcal{C}_t^\gamma \cup \hat{\mathcal{C}}_t^\gamma\right| + \varepsilon}, \tag{17}$$

where ε is a small constant for numerical stability. In the experiments, we report CIoU@γ for $\gamma \in \{100, 90\}$ dB, corresponding to increasingly stricter coverage requirements.

5.3. Quantitative and Qualitative Evaluation

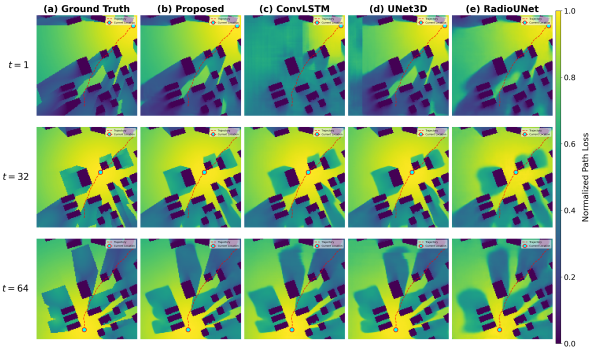

Figure 2. Qualitative comparison on a test environment and the corresponding trajectory: (a) Ground Truth, (b) Proposed, (c) ConvLSTM, (d) 3D U-Net, and (e) RadioUNet. The rows display the radio map at time $t = 1$, 32, and 64, respectively.

Table 1 summarizes both reconstruction accuracy and computational cost. Compared with the strongest baseline 3D U-Net, our model reduces RMSE from 1.0443 to 0.8863 and improves SSIM from 0.9418 to 0.9566. The gain is even more pronounced relative to the frame-wise RadioUNet baseline, which confirms that exploiting inter-frame correlation is important for long-term radio map construction. Importantly, our method also obtains the lowest DM-RMSE, indicating that it better captures frame-to-frame evolution. The consistently higher CIoU values across two thresholds further show that our predictions recover the coverage structure more faithfully.

The lower block of Table 1 shows that our model has a parameter count comparable to ConvLSTM, but requires substantially fewer FLOPs than ConvLSTM, 3D U-Net, and frame-wise RadioUNet. Its inference time is also much lower than sequential ConvLSTM and frame-wise RadioUNet. Although batched RadioUNet achieves the lowest raw inference time, it does not model temporal dependence and therefore exhibits the worst prediction quality. Overall, these results suggest that sharing the scene representation across the trajectory and performing temporal refinement in a compact latent space provides an effective trade-off between accuracy, temporal consistency, and efficiency for long-term radio map generation.

Figure 2 shows a representative qualitative comparison of predicted long-term radio maps, with additional examples provided in Appendix C. Our method produces radio maps that are visually closest to the ground truth. In contrast, the frame-wise RadioUNet baseline tends to produce oversmoothed predictions and less accurate boundary regions. ConvLSTM and 3D U-Net capture some temporal structure, but they can introduce noticeable errors in early frames, which then affect the visual consistency of the sequence.

Table 1. Performance and efficiency on the test set. The upper block reports prediction quality, while the lower block reports model complexity and inference cost. "RadioUNet (batch)" evaluates all frames in parallel and is included only as a throughput-oriented implementation of the frame-wise baseline.

Metric	Ours	ConvLSTM	3D U-Net	RadioUNet	RadioUNet (batch)
RMSE ↓	$\mathbf{0.8863 \pm 0.1892}$	1.3845 ± 0.2587	1.0443 ± 0.2387	1.6942 ± 0.3970	–
SSIM ↑	$\mathbf{0.9566 \pm 0.0157}$	0.9317 ± 0.0219	0.9418 ± 0.0207	0.8953 ± 0.0356	–
DM-RMSE ↓	$\mathbf{0.8547 \pm 0.1892}$	0.9974 ± 0.1988	0.9056 ± 0.1989	1.2773 ± 0.3331	–
CIoU@100 ↑	$\mathbf{0.9438 \pm 0.0271}$	0.9220 ± 0.0357	0.9354 ± 0.0314	0.9131 ± 0.0405	–
CIoU@90 ↑	$\mathbf{0.9674 \pm 0.0160}$	0.9211 ± 0.0233	0.9543 ± 0.0192	0.8503 ± 0.0681	–
Trainable params	7.74M	**7.63M**	10.52M	13.27M	13.27M
FLOPs	**237.67G**	425.24G	494.10G	375.94G	375.94G
Inference time (ms)	22.78 ± 7.43	92.98 ± 0.52	19.96 ± 4.13	175.42 ± 0.18	$\mathbf{14.95 \pm 6.92}$

Table 2. Reduced-frame generation and temporal interpolation for long-term radio map construction. For $K \in \{32, 16, 8\}$, the model is evaluated only at K uniformly sampled anchor frames, and the remaining radio maps are recovered by linear interpolation. The last column evaluates the model at all $T = 64$ frames and serves as the reference.

Metric	$K = 32$	$K = 16$	$K = 8$	Dense ($T = 64$)
RMSE ↓	0.9400 ± 0.1989	1.0860 ± 0.2325	1.4086 ± 0.3059	0.8863 ± 0.1892
SSIM ↑	0.9497 ± 0.0178	0.9331 ± 0.0234	0.9037 ± 0.0336	0.9566 ± 0.0157
DM-RMSE ↓	0.9496 ± 0.2084	1.0291 ± 0.2283	1.0800 ± 0.2401	0.8547 ± 0.1892
CIoU@100 ↑	0.9431 ± 0.0273	0.9410 ± 0.0280	0.9347 ± 0.0308	0.9438 ± 0.0271
CIoU@90 ↑	0.9546 ± 0.0230	0.9312 ± 0.0360	0.8849 ± 0.0604	0.9674 ± 0.0160
FLOPs	119.31G	60.12G	30.53G	237.67G
Inference time (ms)	14.34 ± 0.20	8.79 ± 0.21	6.29 ± 0.22	22.78 ± 7.43

5.4. Reduced-Frame Generation and Temporal Interpolation

This experiment evaluates how generating only a subset of radio map frames affects the reconstruction quality and inference efficiency of long-term radio map. In the dataset, each UAV trajectory is represented by an ordered sequence of $T = 64$ waypoints. The radio map corresponding to the t-th waypoint is treated as the t-th frame of the long-term radio map. To reduce computational cost, we uniformly select $K \in \{32, 16, 8\}$ anchor frame indices from the original sequence, including the first and last frames. The proposed model is evaluated only at these anchor frames to obtain radio map predictions at the associated UAV positions. The complete 64-frame radio map sequence is then reconstructed by applying linear interpolation along the temporal dimension. The interpolated radio map sequences are compared with the ground-truth using the same evaluation protocol as in Section 5.3.

Table 2 reports the interpolation results under different numbers of anchor frames, with dense 64-frame inference included as a reference. When $K = 32$, the proposed scheme preserves 99.28% of the SSIM achieved by dense inference, while reducing FLOPs by $1.99\times$ and inference time by $1.59\times$. The coverage metrics remain nearly unchanged, particularly at $\gamma = 100\,\mathrm{dB}$. When $K = 16$, the reconstructed long-term radio map still retains 97.54% of the dense-inference SSIM, with a $3.95\times$ reduction in FLOPs. Further reducing the number of anchors to $K = 8$ provides the largest computational saving, but causes more significant degradation in RMSE, SSIM, and CIoU. Overall, these results indicate that evaluating the proposed model at a reduced number of anchor frames and interpolating the remaining frames offers a practical accuracy-efficiency trade-off for long-term radio map construction.

6. Conclusion

We studied long-term radio map construction in a UAV-assisted communication scenario, where the transmitter moves along a known UAV trajectory over a static environment. Different from the conventional static radio map estimators, we formulated the problem as a conditioned video generation task, enabling the model to fully capture the spatiotemporal correlations among consecutive radio map frames. The proposed model combined shared scene encoding, per-frame transmitter conditioning, and latent temporal refinement, which together provided a principled way to balance spatial fidelity, temporal coherence, and computational efficiency. An important direction for future work is to consider more challenging scenarios, such as online causal prediction, and joint integration with downstream tasks such as trajectory planning, handover control, and resource allocation.

References

Bi, S., Lyu, J., Ding, Z., and Zhang, R. Engineering radio maps for wireless resource management. *IEEE Transactions on Wireless Communications*, 26(2):133–141, April 2019.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., and Rombach, R. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Cheng, N., Jia, H., Wang, X., Peng, H., Sun, R., and Zhou, C. Radiomapmotion: A dataset and benchmark for proactive spatio-temporal radio environment prediction. *IEEE Transactions on Cognitive Communications and Networking*, pp. 1–14, 2026.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

Hoffmann, M. and Kryszkiewicz, P. Beam management driven by radio environment maps in O-RAN architecture. In *Proceedings of the 2023 IEEE International Conference on Communications Workshops*, pp. 54–59, 2023.

Hoppe, R., Wölfle, G., and Jakobus, U. Wave propagation and radio network planning software WinProp added to the electromagnetic solver package FEKO. In *Proceedings of the 2017 International Applied Computational Electromagnetics Society Symposium*, pp. 1–2, 2017.

Levie, R., Yapar, Ç., Kutyniok, G., and Caire, G. RadioUNet: Fast radio map estimation with convolutional neural networks. *IEEE Transaction on Wireless Communications*, 20(6):4001–4015, June 2021.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. C. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018.

Romero, D. and Kim, S.-J. Radio map estimation: A data-driven approach to spectrum cartography. *IEEE Signal Processing Magazine*, 39(6):53–72, November 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, 2015.

Wang, X., Tao, K., Cheng, N., Yin, Z., Li, Z., Zhang, Y., and Shen, X. RadioDiff: An effective generative diffusion model for sampling-free dynamic radio map construction. *IEEE Transactions on Cognitive Communications and Networking*, 11(2):738–750, April 2025.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Wu, Q., Zeng, Y., and Zhang, R. Joint trajectory and communication design for multi-UAV enabled wireless networks. *IEEE Transactions on Wireless Communications*, 17(3): 2109–2121, 2018.

Yapar, Ç., Levie, R., Kutyniok, G., and Caire, G. Real-time outdoor localization using radio maps: A deep learning approach. *IEEE Transaction on Wireless Communications*, 22(12):9703–9717, December 2023.

Zeng, Y., Wu, Q., and Zhang, R. Accessing from the sky: A tutorial on UAV communications for 5G and beyond. *Proceedings of the IEEE*, 107(12):2327–2375, 2019.

Zeng, Y., Chen, J., Xu, J., Wu, D., Xu, X., Jin, S., Gao, X., Gesbert, D., Cui, S., and Zhang, R. A tutorial on environment-aware communications via channel knowledge map for 6G. *IEEE Communications Surveys & Tutorials*, 26(3):1478–1519, 2024.

Zhang, S. and Zhang, R. Radio map-based 3D path planning for cellular-connected UAV. *IEEE Transaction on Wireless Communications*, 20(3):1975–1989, March 2021.

Zhang, S., Zeng, Y., and Zhang, R. Cellular-enabled UAV communication: A connectivity-constrained trajectory optimization perspective. *IEEE Transactions on Wireless Communications*, 67(3):2580–2604, March 2019.

Zhang, S., Wijesinghe, A., and Ding, Z. RME-GAN: A learning framework for radio map estimation based on conditional generative adversarial network. *IEEE Internet of Things Journal*, 10(20):18016–18027, October 2023.

Zhang, S., Choi, B., Ouyang, F., and Ding, Z. Physics-inspired machine learning for radiomap estimation: Integration of radio propagation models and artificial intelligence. *IEEE Communications Magazine*, 62(8):155–161, August 2024.

Zhu, L., Zhu, W., Zhang, S., Caire, G., and Liu, L. DF-3DRME: A data-friendly learning framework for 3D radio map estimation based on super-resolution technique. *arXiv preprint arXiv:2604.00676*, 2026a.

Zhu, L., Zhu, W., Zhang, S., Caire, G., and Liu, L. A data-friendly deep learning architecture for high-resolution radio map construction. In *Proceedings of IEEE International Conference on Communications Workshops*, pp. 1–6, May 2026b.

Zhu, W., Tao, M., Yuan, X., and Guan, Y. Message passing-based joint user activity detection and channel estimation for temporally-correlated massive access. *IEEE Transactions on Communications*, 71(6):3576–3591, Jun. 2023.

A. Detailed Model Architecture

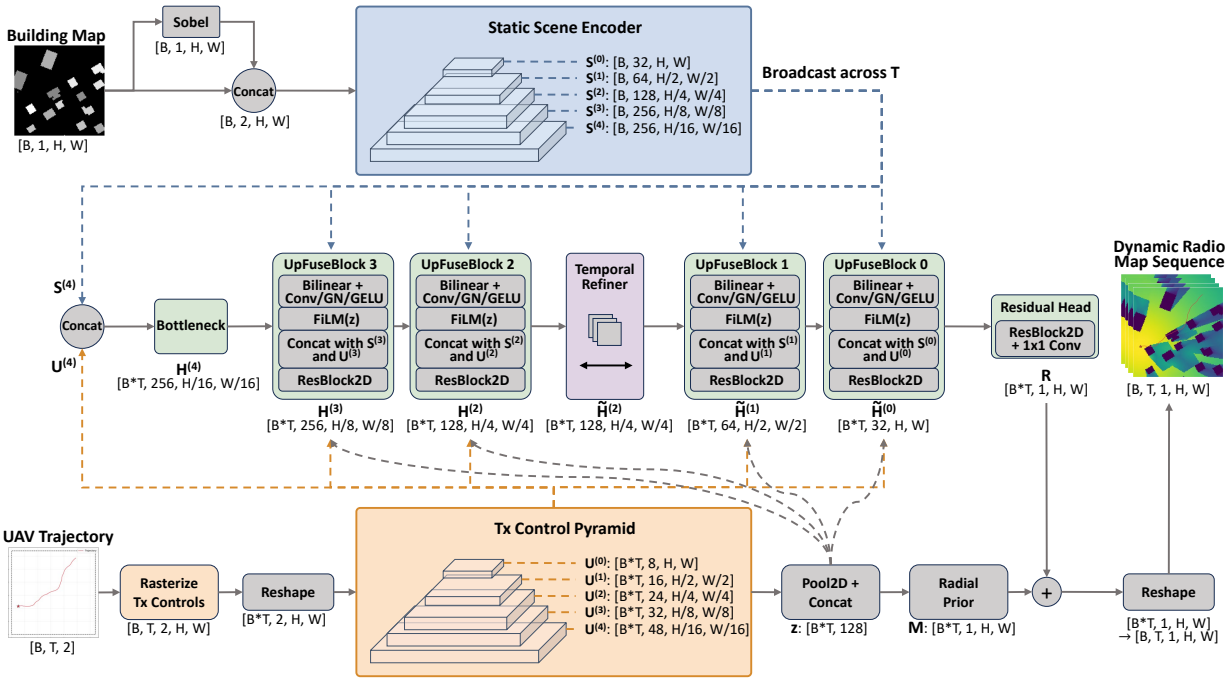

Figure 3. Detailed architecture of the proposed model for long-term radio map construction. The figure illustrates the main processing modules, information flow, and intermediate representations used to generate long-term radio map from the environment representation and UAV trajectory.

Figure 3 provides the detailed module-level architecture of the proposed model, complementing the high-level overview presented in the main text.

B. Long-term Radio Map Dataset

We construct a long-term radio map dataset using a ray-tracing simulator (Hoppe et al., 2017) over 320 urban environments extracted from OpenStreetMap. The environments are sampled from four cities: Ankara, Berlin, Ljubljana, and Tel Aviv. Each environment covers a $128\,\mathrm{m} \times 128\,\mathrm{m}$ area and is discretized with a spatial resolution of $\Delta = 1\,\mathrm{m}$, resulting in an environment representation $\mathbf{E} \in \mathbb{R}^{128\times128}$.

For each environment, we generate 30 UAV trajectories. Each trajectory is represented by an ordered sequence of $T = 64$ waypoints, with a spacing of $2\,\mathrm{m}$ between adjacent waypoints. This results in a total of $N = 9{,}600$ long-term radio maps, corresponding to $614{,}400$ individual radio map snapshots.

We split the dataset at the *environment* level: 70% of the environments, together with all associated trajectories, are used for training, 20% for validation, and 10% for testing. This split prevents building layouts from appearing in multiple partitions and therefore provides a stricter evaluation of generalization to unseen urban environments.

C. Additional Qualitative Results

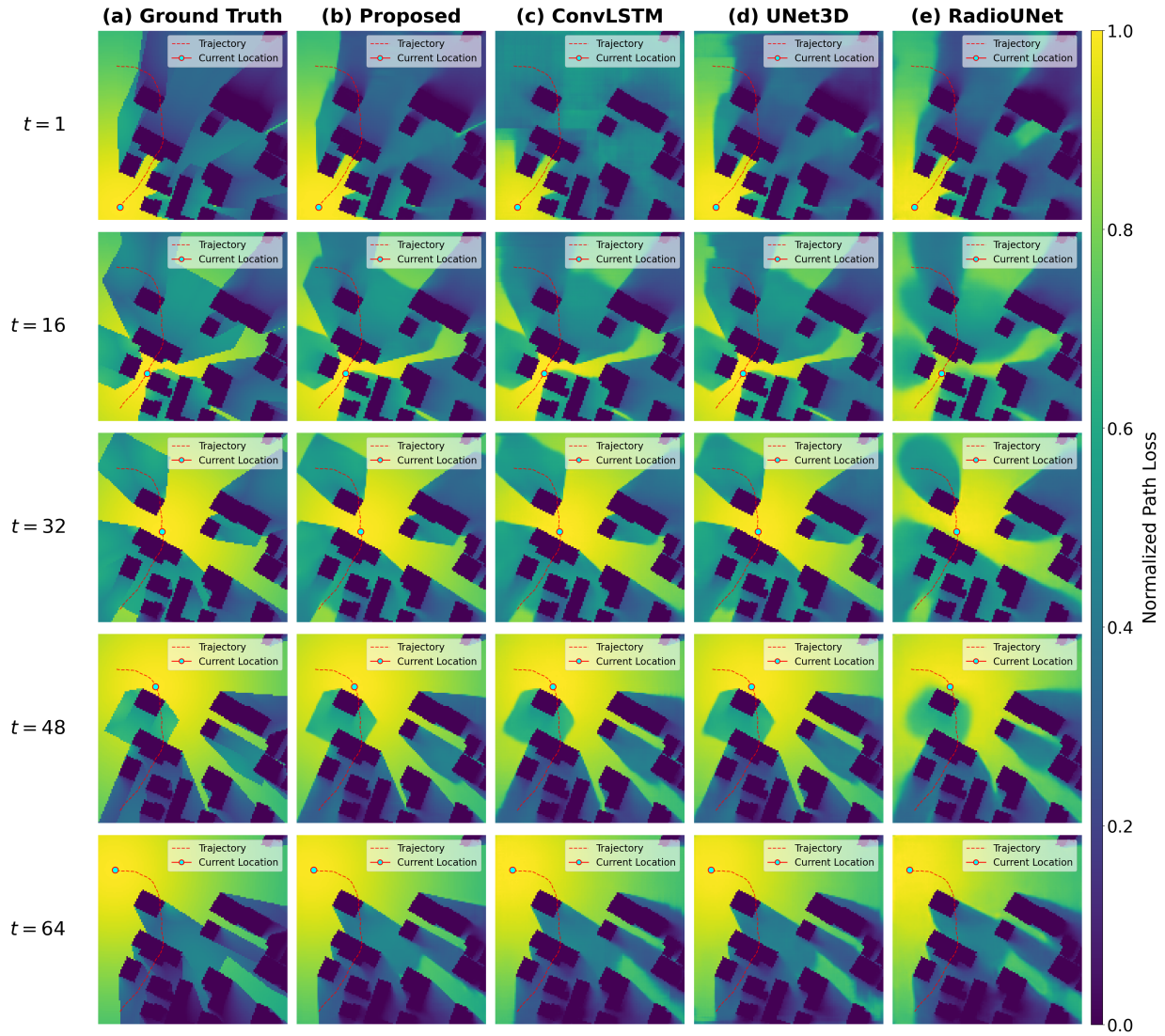

Figure 4. Qualitative comparison on a test environment and the corresponding trajectory: (a) Ground Truth, (b) Proposed, (c) ConvLSTM, (d) 3D U-Net, and (e) RadioUNet. The rows display the radio map at time $t = 1$, 16, 32, 48, and 64, respectively.