SRDiff: A Cross-Modal Diffusion Model for Satellite-to-Radar Translation in Precipitation Nowcasting

You Qin, Jinming Cao, Ting Wang, Yifang Yin, Li Li, Shili Xiang, Ying Zhang, Roger Zimmermann *Senior Member, IEEE*

***Abstract*—Satellite-to-radar translation is a critical yet under-explored task in modern weather forecasting. While satellites provide near-global coverage and high-frequency updates, their multi-channel radiance data is difficult to interpret directly by forecasters or end-users, and not readily compatible with existing radar-based forecasting models. By converting satellite observations into radar-equivalent representations, we can (i) extend radar-like availability to regions without ground-based radar coverage, (ii) provide a more compact and operationally meaningful representation of storms, and (iii) enable the reuse of a large ecosystem of radar-based nowcasting models without retraining. Despite its importance, methods tailored for this task remain limited, and the performance of general-purpose generative models has not been systematically benchmarked. To fill this gap, we propose SRDiff, a cross-modal, sequence-aware diffusion model specifically designed for satellite-to-radar translation. SRDiff employs a Cross-Modal Conditional Adapter to align heterogeneous satellite channels with radar reflectivity, and a Diffusion Transformer backbone to capture sequential dependencies, producing stable and coherent radar predictions. Extensive experiments on the SEVIR and Sat2Rdr datasets show that SRDiff outperforms existing deterministic and diffusion-based methods on the satellite-to-radar translation task across all key metrics. As a downstream application, we integrate SRDiff with off-the-shelf radar-based nowcasting models, demonstrating three key advantages: (i) operational efficiency, since no new training is required, (ii) strong forecasting accuracy, often surpassing models trained from scratch, and (iii) practical validation that effective satellite-to-radar translation directly improves downstream nowcasting. Code is available at https://github.com/42xingxing/SRDiff.**

***Index Terms*—Satellite-to-radar translation, precipitation nowcasting, diffusion models, cross-modal learning, extreme weather forecasting**

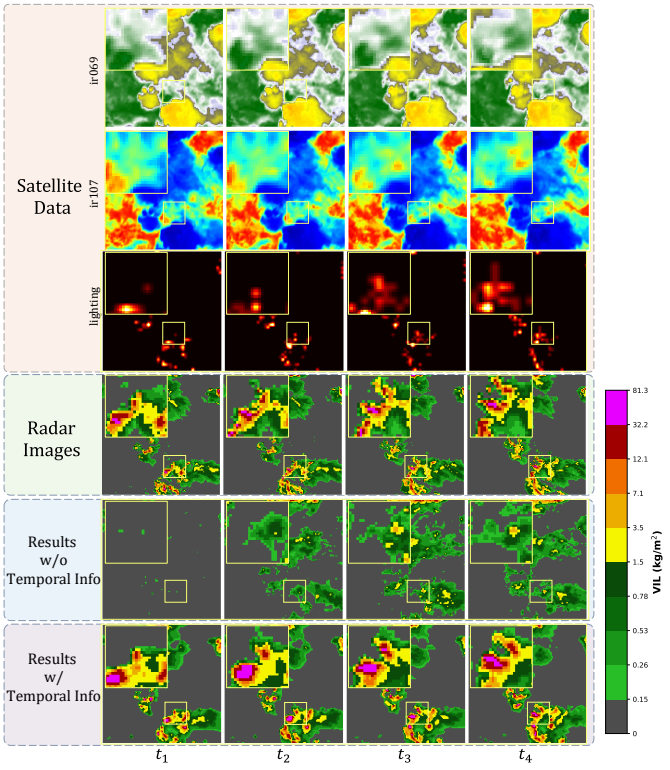

Fig. 1. Challenges in Satellite-to-Radar Translation. (a) Modality Gap: The first four rows show satellite inputs (IR069, IR107, lightning) and the corresponding radar-derived Vertically Integrated Liquid (VIL) fields, highlighting the lack of direct mapping. (b) Temporal Dependency: The last two rows compare predictions without and with temporal modeling—temporal context reduces fluctuations and improves consistency.

I. Introduction

Weather radar reflectivity is the cornerstone of modern short-term precipitation forecasting. It provides high-resolution depictions of storm structures, rainfall intensities, and convective evolution, which are indispensable for disaster mitigation, air traffic scheduling, hydropower management, and urban drainage planning [1], [2], [3]. Numerous advanced nowcasting methods [4], [5], [6], [7], [8], [9], [10], [11], [12] have been built upon radar sequences, achieving remarkable success in predicting rainfall up to a few hours ahead. Recent advances have further extended deep learning to broader weather forecasting tasks, including tropical cyclone track and intensity prediction using large-scale benchmark datasets [13], as well as multi-modal diffusion models for tropical cyclone precipitation forecasting [14], demonstrating the growing versatility of data-driven approaches in operational meteorology. However, radar coverage is geographically sparse and uneven [15]: many developing regions, oceans, and mountainous areas lack dense radar networks, leading to blind spots in precipitation monitoring. Even in highly urbanized regions, such as Singapore, radar observations remain spatially constrained. Singapore's weather radar network primarily covers the main island and nearby surrounding areas, but observations over adjacent maritime

You Qin, Jinming Cao, Roger Zimmermann are with National University of Singapore, Singapore (qinyou@u.nus.edu, jinming.ccao@gmail.com, rogerz@comp.nus.edu.sg)

Ting Wang and Ying Zhang are with the School of Computer Science, Northwestern Polytechnical University, Xi'an, China (tingw@mail.nwpu.edu.cn, yingz118@gmail.com)

Yifang Yin and Shili Xiang are with Institute for Infocomm Research (I^2R), A*STAR (yin_yifang@a-star.edu.sg, sxiang@a-star.edu.sg)

Li Li is with the University of Southern California, Los Angeles, CA, USA (li.li02@usc.edu)

Corresponding authors: jinming.ccao@gmail.com, yin_yifang@i2r.a-star.edu.sg

TABLE I
COMPARISON OF QPE PRODUCTS AND THE PROPOSED SATELLITE-TO-RADAR TRANSLATION APPROACH. TRADITIONAL QPE PRODUCTS ESTIMATE PRECIPITATION RATES WITH SIGNIFICANT LATENCY, WHILE OUR SATELLITE-TO-RADAR TRANSLATION METHOD GENERATES RADAR-DERIVED PRODUCTS IN REAL TIME FOR DEEP LEARNING-BASED NOWCASTING.

Product (Dataset)	Primary Source	Spatial	Temporal	Latency	Output	Application
Traditional QPE Products						
IMERG Early	PMW (LEO)	0.1° (∼10 km)	30 min	4 hours	mm/hr	Rapid-response monitoring
IMERG Late	PMW (LEO)	0.1° (∼10 km)	30 min	14 hours	mm/hr	Next-day monitoring
IMERG Final	PMW (LEO)	0.1° (∼10 km)	30 min	3.5 months	mm/hr	Climate research
CMORPH	PMW (LEO)	8 km	30 min	∼3 hours	mm/hr	Hydrology applications
PERSIANN-CCS	IR (GEO)	0.04° (∼4 km)	1 hour	∼1 hour	mm/hr	Real-time monitoring
Satellite-to-Radar Translation (This Work)						
SRDiff (SEVIR)	GOES-16 (GEO)	**1 km**	**5 min**	**Real-time**	**VIL (kg/m^2)**	**DL nowcasting**
SRDiff (Sat2Rdr)	GK2A (GEO)	**2 km**	**10 min**[1]	**Real-time**	**HSR (mm/hr)**	**DL nowcasting**

regions, where many convective systems originate, are limited by range, beam blockage, and attenuation effects in intense tropical rainfall. As a result, precipitation systems approaching from the surrounding seas may not be fully captured until they move within effective radar range, reducing the available lead time for accurate nowcasting. This observation highlights that reliance on radar-only forecasting may be insufficient for seamless regional precipitation monitoring and motivates the integration of satellite data to fill observational gaps.

A long line of research has addressed satellite-based precipitation estimation through Quantitative Precipitation Estimation (QPE) products such as IMERG [16], CMORPH [17], PERSIANN [18], and GPM-based retrievals [19]. These products have been highly successful for hydrological modeling, climate research, and flood monitoring. However, QPE and satellite-to-radar translation address fundamentally different needs. QPE products estimate *precipitation rates* (mm/hr) and are optimized for physical retrieval accuracy. Comparatively, our goal is to generate *radar reflectivity* or its derived products such as Vertically Integrated Liquid (VIL) or Hybrid Surface Rainfall (HSR), which preserve the spatiotemporal structure required by deep learning-based nowcasting models. The two approaches are complementary: QPE serves applications where accurate rainfall amounts are critical, while satellite-to-radar translation enables the reuse of pre-trained radar nowcasting models in regions lacking ground radar infrastructure.

Beyond task formulation, QPE products and satellite-to-radar translation differ substantially in spatial resolution, temporal sampling, and operational latency. As summarized in Table I, traditional QPE products such as IMERG and CMORPH primarily rely on low-Earth orbit (LEO) passive microwave (PMW) sensors with coarse spatial resolution (∼10 km), temporal intervals of 30 min or longer, and significant processing latency ranging from hours to months. Comparatively, our proposed satellite-to-radar translation model leverages geostationary (GEO) platforms, *e.g.*, GOES-16 or GK2A, which provide continuous observations at 1-2 km resolution and 5-10 min cadence with real-time availability. For deep learning-based nowcasting at sub-hourly scales, such continuous high-resolution coverage is essential. SRDiff is designed to leverage this advantage and produce translation results at the data resolution in real time.

As discussed above, meteorological satellites offer near-global coverage and frequent updates [20], [21], [22]. With multiple infrared, visible, and lightning channels, they provide rich information on cloud morphology, thermal profiles, and storm electrification processes. Yet, these advantages also bring challenges. Satellite observations are multi-channel, heterogeneous, and often physically indirect with respect to precipitation intensity. For forecasters and end-users, raw satellite imagery can be overwhelming. Meanwhile, for existing machine learning models designed for radar data, satellite observations are incompatible in both representation and resolution. Thus, the problem of satellite-to-radar translation emerges as a critical step, enabling the generation of radar-equivalent representations from satellite observations.

The significance of this translation extends beyond simple data conversion. First, it allows the vast ecosystem of radar-based nowcasting models to be reused without modification, providing immediate operational benefits. Second, radar reflectivity provides a standardized and compact representation. Compared with the numerous channels in satellite data, radar reflectivity and its derived VIL/HSR fields are more intuitive to interpret for both forecasters and automated downstream systems. Third, synthesizing radar-like data from satellites effectively extends radar availability to regions without ground-based coverage, making radar-based forecasting globally accessible. In addition, generating radar-equivalent outputs provides a common representation that bridges short-term nowcasting and longer-range numerical weather prediction (NWP) forecasts, allowing seamless model blending within hybrid physics-AI forecasting pipelines.

Despite its importance, there are currently no models explicitly designed for satellite-to-radar translation. To provide a baseline comparison, we adopt representative image-to-image translation approaches such as Pix2Pix, CycleGAN, and StyleGAN [23], [24], [25], which are natural candidates for cross-modal mapping but not tailored to this domain. These generic methods struggle with the modality gap between heterogeneous sensing systems and often fail to capture the

[1]The paired satellite and radar data are further downsampled to a temporal resolution of 1 hour in the Sat2Rdr dataset.

temporal dependencies critical for precipitation dynamics [26], [27]. As a result, their generated radar fields tend to suffer from instability across time and lack physical consistency.

The following two challenges make the task particularly demanding. First, the modality gap: as illustrated in Fig. 1(a), satellite radiances (*e.g.*, IR069, IR107, lightning) and radar-derived VIL capture different physical aspects of weather systems. Even structurally similar satellite inputs may correspond to distinct radar outputs, making direct one-to-one mapping infeasible [28]. Second, the temporal dependency: precipitation is inherently dynamic, and radar fields evolve according to atmospheric processes that satellites only indirectly observe. Ignoring sequential context, as shown in Fig. 1(b), leads to unstable predictions with frame-to-frame fluctuations, undermining their forecasting utility.

To address these gaps, we introduce SRDiff, a cross-modal, sequence-aware diffusion model explicitly designed for satellite-to-radar translation. Unlike previous work that adopts general-purpose architectures [21], [22], [27], SRDiff is tailored to the task in two key ways. First, a Cross-Modal Conditional Adapter (CMCA) refines and aligns heterogeneous satellite inputs into a compact feature space compatible with radar reflectivity. Second, a Diffusion Transformer (DiT) backbone captures long-range temporal dependencies, ensuring coherent and physically plausible radar sequences. Together, these components bridge the sensing modalities and temporal dynamics, yielding stable high-quality radar predictions.

Beyond the translation task itself, we further demonstrate the practical utility of SRDiff through the downstream application of precipitation nowcasting. Instead of training new models on satellite data, we directly plug converted radar outputs into existing radar-based forecasting systems [8], [4], [29], [7], [30], [31], [32], [33], [34]. This offers three key advantages. (i) Operational efficiency: forecasters can deploy SRDiff as a preprocessing module, avoiding the need to retrain or redesign existing nowcasting models. (ii) Forecasting accuracy: experiments show that SRDiff-enhanced inputs achieve results comparable to, or better than, models trained from scratch on satellite-radar joint data. (iii) Validation by downstream performance: the improvements in translation quality translate directly into superior forecasting results, confirming the practical value of our approach.

We evaluate SRDiff extensively on two public benchmark datasets, namely SEVIR [35] and Sat2Rdr [28]. SRDiff consistently surpasses deterministic baselines [23], [24], [26], [25] and recent diffusion-based approaches [29], [22] across multiple metrics. Furthermore, in downstream nowcasting experiments, SRDiff achieves leading performance on critical success indices (CSI-M, CSI-181, CSI-219), highlighting its robustness in capturing both widespread precipitation and extreme convective events. In summary, this work makes the following key contributions:

- We formally establish satellite-to-radar translation as an operationally significant but underexplored task, bridging the gap between globally accessible satellite observations and regional radar-based forecasting systems.
- We propose SRDiff, which, to the best of our knowledge, is the first cross-modal, sequence-aware diffusion model explicitly designed for this task. It introduces CMCA and DiT to jointly address the modality gap and temporal dependency.
- We conduct comprehensive benchmarking of satellite-to-radar translation on SEVIR and Sat2Rdr, demonstrating that SRDiff achieves state-of-the-art performance against deterministic and diffusion-based baselines.
- We validate the operational utility of SRDiff by integrating it with existing radar-based nowcasting systems, where it consistently improves downstream forecasting skills, particularly on critical success indices (CSI-M, CSI-181, and CSI-219).

II. Related Work

A. *Satellite Precipitation Estimation*

Satellite-based precipitation estimation has been an active research area for decades. Passive microwave (PMW) sensors aboard low-Earth orbit (LEO) satellites, beginning with TRMM [36] and continuing with the GPM Core Observatory [19], exploit the direct physical interaction between microwave radiation and hydrometeors (emission over oceans, scattering over land) to retrieve instantaneous rain rates with high accuracy. The GPM Dual-frequency Precipitation Radar (DPR) provides the most accurate spaceborne precipitation measurements and serves as calibration reference for the global PMW constellation. Multi-satellite products such as IMERG [16], CMORPH [17], and PERSIANN [18] merge PMW retrievals with infrared-based estimates to achieve near-global coverage at 0.1°/30 min resolution. These products have proven invaluable for hydrological modeling, climate research, and flood monitoring. However, despite the physical advantages of PMW sensing, LEO orbits impose a fundamental temporal limitation: each satellite provides only 2–4 overpasses per day over a given location. Even with the full GPM constellation ($\sim$10 PMW radiometers), minute-scale temporal continuity cannot be guaranteed. For precipitation nowcasting, which requires tracking storm evolution at 5–10 min intervals, such sparse sampling is often insufficient, and the hour- to month-scale latency of these products further constrains their applicability to real-time nowcasting.

Comparatively, our method SRDiff takes a complementary approach by leveraging geostationary satellite observations (GOES-16, GK2A), which provide continuous imaging at 5 min intervals. While IR and water vapor channels lack the physical directness of PMW for precipitation retrieval, SRDiff learns to bridge this gap through data-driven training, generating radar-equivalent fields that preserve the spatiotemporal structure required by deep learning-based nowcasting models. The two paradigms serve different purposes: PMW/QPE products excel at instantaneous precipitation accuracy for climate and hydrology applications, while SRDiff enables real-time radar-format sequence generation for integration with the downstream nowcasting systems.

B. *Precipitation Nowcasting with Deep Learning*

Precipitation nowcasting is vital for extreme weather preparedness and hydrological management, especially under

accelerating climate change. Deep learning approaches can be broadly divided into deterministic and probabilistic models. Deterministic models have advanced via architectural innovations: ConvLSTM [4] introduced spatiotemporal modeling, PredRNN [37] improved memory with decoupled cell states, and PhyDNet [38] incorporated physics-guided motion consistency. Later models such as Earthformer [5], SimVP [34], and spacetime cuboid attention introduced non-recurrent, efficient designs. However, L2-based losses often yield oversmoothed predictions.

Probabilistic models aim to capture uncertainty. GAN-based approaches like DGMR [39] and NowcastNet [40] enhance realism but suffer from training instability. Diffusion models offer a more stable alternative: PreDiff [41] introduces physics-conditioned latent spaces, and Cascast [29] applies cascaded refinement for realism. DiffCast [42] proposes a unified residual diffusion framework that decomposes precipitation fields into deterministic global motion and stochastic local variations, achieving state-of-the-art results on multiple radar benchmarks. Wang *et al.* [43] incorporate physical constraints into diffusion networks for skilful radar-based nowcasting. RNDiff [44] introduces a condition diffusion model that leverages auxiliary meteorological fields to improve rainfall prediction quality. Pan *et al.* [45] propose a short-long term sequence learning network that captures multi-scale temporal dependencies for precipitation nowcasting.

While effective with radar data, these methods are limited in radar-sparse regions, a gap that the proposed SRDiff framework aims to address. Satellite-conditioned models often use basic UNet [26] or Transformer architectures [46], directly ingesting raw satellite data without modality alignment. This undermines accuracy, as satellite signals require structured translation to reflect precipitation effectively. Recent satellite-only models like NPM [28] also lack ground-truth guidance and need post-hoc radar translation. Our work addresses these limitations via an explicit satellite-to-radar translation, enhancing both supervision and downstream forecasting accuracy.

C. Translation Models

Modern image translation approaches fall into three categories: (1) UNet-based models with skip connections for structural preservation, (2) GAN-based methods such as Pix2Pix [23] (paired) and CycleGAN [24]/UVCGAN v2 [47] (unpaired), and (3) diffusion models (*e.g.*, Palette [48], BBDM [49]) that refine outputs iteratively. While StegoGAN [25] improves semantic consistency via steganography and VideoFusion [50] handles temporal coherence, these methods assume single-modality inputs with structural alignment. Our work addresses the more challenging cross-modal translation from satellite to radar, where sensing physics differs and direct pixel-level correspondence is absent, necessitating new solutions to ensure both temporal coherence and physical consistency.

III. SRDiff

A. Overall Architecture and Notations.

Building upon the identified challenges in satellite-to-radar translation, namely the modality gap between satellite radiance and radar reflectivity, as well as the critical role of temporal dependencies, we propose SRDiff, a diffusion-based satellite-to-radar sequence translation model. As illustrated in Fig. 2, SRDiff receives a sequence of multi-modal satellite observations $\mathbf{X}^c_{0:T} \in \mathbb{R}^{T\times H_c\times W_c}$, where $c \in \mathcal{C}$ denotes the set of input modalities. For each dataset we consider, $\mathcal{C}$ includes all relevant satellite channels publicly available in that dataset. For example, $\mathcal{C} = \{\texttt{ir107}, \texttt{ir069}, \texttt{lght}\}$ for the SEVIR (GOES-16) dataset or $\mathcal{C} = \{\texttt{ir105}, \texttt{wv063}, \texttt{wv073}\}$ for the Sat2Rdr (GK2A) dataset, representing different observation modalities over the past T time steps. The objective is to translate these satellite observations into the corresponding radar reflectivity fields (or their derived products, depending on the dataset) $x_{0:T} \in \mathbb{R}^{T\times H\times W}$, where H, W and H_c, W_c denote the spatial resolutions of radar and satellite data, respectively, covering the same geographic region.

Specifically, SRDiff consists of two key components: (1) the **Cross-Modal Conditional Adapter (CMCA)**, which extracts and fuses informative features from multi-modal satellite inputs, providing adaptive conditioning for the generative process; and (2) the **Satellite-to-Radar Diffusion Model**, which utilizes a spatiotemporal DiT as the denoising backbone to iteratively refine high-fidelity radar reflectivity from its satellite-derived counterparts. By combining the powerful generative capabilities of diffusion models with CMCA's cross-modal feature alignment, SRDiff effectively bridges the modality gap while preserving temporal coherence in radar predictions. This framework enhances precipitation nowcasting accuracy and generalizability, particularly in regions without direct radar coverage.

B. Cross-Modal Conditional Adapter (CMCA)

The CMCA module integrates multi-modal satellite features through a primary–auxiliary modality framework. We designate the satellite channels most directly correlated with radar observations (*e.g.*, infrared channels capturing cloud-top structure) as the *primary modality*, and channels providing complementary but sparser or indirect cues (*e.g.*, lightning or water vapor) as the *auxiliary modality*. This distinction is motivated by the observation that auxiliary signals, although informative, cannot independently serve as evidence for radar translation due to their sparse or indirect nature. Accordingly, CMCA employs two separate modal-specific encoders: one for the primary modality and one for the auxiliary, followed by a Modality-Temporal Attention Integrator (MTAI) and a Post-Fusion Encoder. This two-encoder design naturally accommodates different numbers of channels per modality (*e.g.*, 2 IR channels + 1 lightning for SEVIR; 1 IR + 2 water vapor for Sat2Rdr), while the total number of encoders remains fixed at two regardless of channel count. In the following, we use SEVIR modality names (IR for primary, Lightning for auxiliary) for concreteness; the formulation readily generalizes to other modality pairs.

Modal-specific Encoder. The primary and auxiliary modalities are encoded by separate 3D convolutional encoders sharing the same architecture but with independent parameters. Given primary input $\mathbf{X}^{IR}_{0:T} \in \mathbb{R}^{T\times C_p\times H_c\times W_c}$ (where C_p denotes

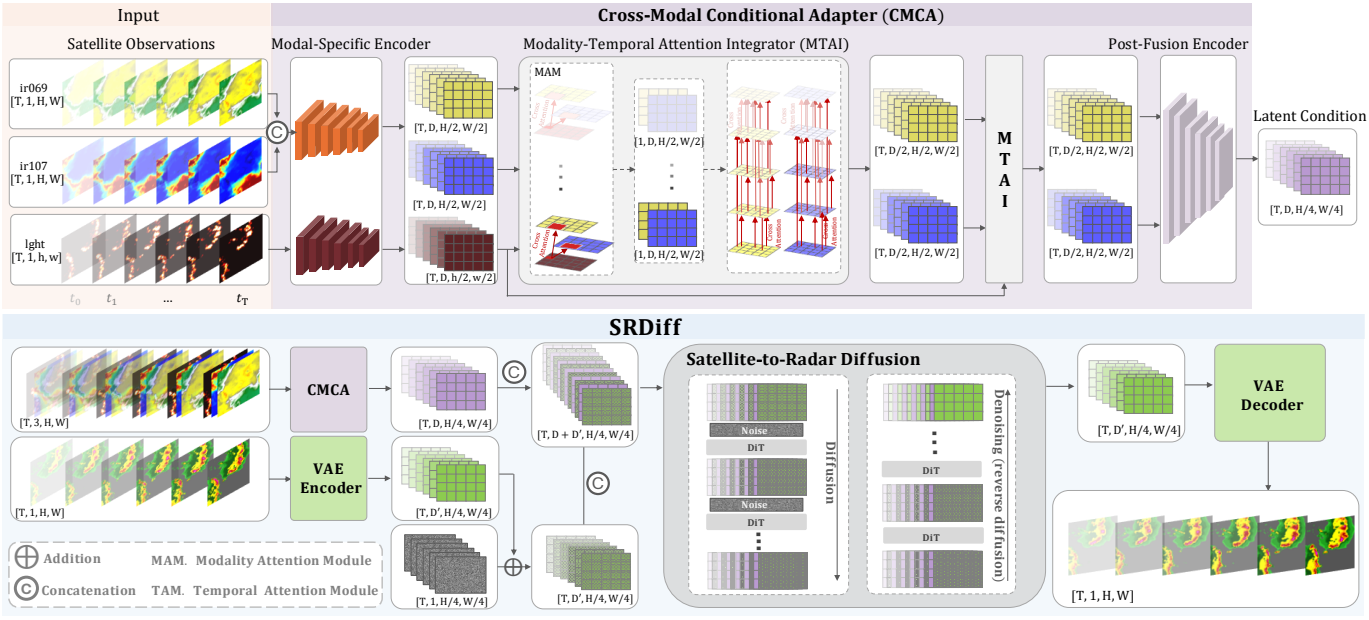

Fig. 2. Overview of SRDiff. The upper part illustrates the Cross-Modal Conditional Adapter (CMCA), which processes multi-channel satellite observations to generate latent conditions. The lower part shows the full SRDiff pipeline, consisting of the CMCA module and a Satellite-to-Radar Diffusion Model, enabling translation of radar-like outputs from satellite data.

the number of primary channels, *e.g.*, $C_p = 2$ for ir069 and ir107) and auxiliary input $\mathbf{X}^L_{0:T} \in \mathbb{R}^{T\times C_a\times H_c\times W_c}$ ($C_a = 1$ for lightning), each encoder applies a 3D convolution with temporal stride 1 and spatial stride 2, followed by two 3D residual blocks, producing:

$$\mathbf{F}^{IR}_{0:T} \in \mathbb{R}^{T\times d\times H'\times W'}, \quad \mathbf{F}^{L}_{0:T} \in \mathbb{R}^{T\times d\times H'\times W'},$$

where $H' = H_c/2$, $W' = W_c/2$, and d is the shared embedding dimension. Only the initial convolution differs in its input channel count (C_p vs. C_a); all subsequent attention and diffusion layers remain identical, making the architecture inherently channel-agnostic.

Modality-Temporal Attention Integrator. To achieve effective feature fusion across modalities and time, MTAI consists of two core components: 1. Modality Attention Module (MAM) for cross-modal interaction. 2. Temporal Attention Module (TAM) for enhancing temporal coherence. MTAI is a repeatable module, allowing flexible iterations based on computational constraints. In this work, we apply MTAI twice to progressively refine the fused features. In each iteration, MAM computes cross-modal attention between the (iteratively updated) primary features and the fixed auxiliary features, followed by TAM for temporal fusion.

(a) Modality Attention Module: MAM performs cross-modal feature alignment via bidirectional attention, following the BiDAF framework [51]. For each time step independently, both feature maps are first transformed into patch token sequences via patch embedding with patch size p:

$$\mathbf{F}^{IR}_{0:T} \in \mathbb{R}^{T\times M\times d}, \quad \mathbf{F}^{L}_{0:T} \in \mathbb{R}^{T\times M\times d},$$

where $M = H'W'/p^2$ is the number of patches per frame. Both modalities share the same spatial dimensions after encoding, so the token count M is identical.

A per-frame cross-modal similarity matrix $\mathcal{S}_{0:T} \in \mathbb{R}^{T\times M\times M}$ is computed via a learnable trilinear function [51]. For frame t, the (i, j)-th entry measures the relevance between primary token i and auxiliary token j:

$$\mathcal{S}_{t,ij} = \mathbf{w}_C^\top \mathbf{f}^{IR}_{t,i} + \mathbf{w}_Q^\top \mathbf{f}^{L}_{t,j} + \left(\mathbf{f}^{IR}_{t,i} \odot \mathbf{w}_m\right)^\top \mathbf{f}^{L}_{t,j}, \tag{1}$$

where $\mathbf{w}_C, \mathbf{w}_Q, \mathbf{w}_m \in \mathbb{R}^d$ are learnable parameters capturing independent primary contributions, independent auxiliary contributions, and their multiplicative interaction, respectively.

Two normalized versions are derived: $\mathcal{S}_r = \mathrm{softmax}_{\mathrm{row}}(\mathcal{S})$ (normalized along the auxiliary dimension, each row sums to 1) and $\mathcal{S}_c = \mathrm{softmax}_{\mathrm{col}}(\mathcal{S})$ (normalized along the primary dimension, each column sums to 1). These yield two complementary cross-modal attention representations:

$$\begin{aligned} \mathcal{A}_{0:T} &= \mathcal{S}_r \cdot \mathbf{F}^{L}_{0:T} \in \mathbb{R}^{T\times M\times d}, \\ \mathcal{B}_{0:T} &= \mathcal{S}_r \cdot \mathcal{S}_c^\top \cdot \mathbf{F}^{IR}_{0:T} \in \mathbb{R}^{T\times M\times d}. \end{aligned} \tag{2}$$

Here, $\mathcal{A}$ (*single-hop attention*) computes, for each primary token, a weighted sum of auxiliary features based on their cross-modal relevance, representing auxiliary information most pertinent to each primary spatial location. $\mathcal{B}$ (*double-hop attention*) re-routes primary features through the auxiliary modality via $\mathcal{S}_c^\top$ and back via $\mathcal{S}_r$, capturing second-order cross-modal dependencies. The double-hop is particularly important for sparse auxiliary signals (*e.g.*, lightning): since auxiliary observations alone are too sparse for dense prediction, $\mathcal{B}$ propagates cross-modal correlations to spatial regions lacking direct auxiliary evidence by leveraging the richer primary feature space.

The final multi-modal representation fuses four complementary views:

$$\mathbf{F}^{M}_{0:T} = \mathtt{FFN}\left([\mathbf{F}^{IR}_{0:T};\ \mathcal{A}_{0:T};\ \mathbf{F}^{IR}_{0:T} \odot \mathcal{A}_{0:T};\ \mathbf{F}^{IR}_{0:T} \odot \mathcal{B}_{0:T}]\right), \tag{3}$$

where $[\cdot;\cdot]$ denotes concatenation along the feature dimension, $\odot$ is element-wise multiplication, and FFN is a linear projection from $4d$ to d. These four terms respectively provide: (i) original primary features preserving the baseline spatial structure, (ii) attended auxiliary features representing direct cross-modal evidence, (iii) element-wise co-activation highlighting where primary and auxiliary signals spatially co-occur, and (iv) primary features modulated by bidirectional attention capturing indirect cross-modal relationships. This multi-view fusion ensures comprehensive information exchange between the two modalities.

(b) Temporal Attention Module: After unpatchification restores the spatial dimensions, TAM enhances temporal coherence by applying multi-head self-attention independently at each spatial location along the time axis. Concretely, the feature map $\mathbf{F}^M_{0:T} \in \mathbb{R}^{T\times d\times H'\times W'}$ is reshaped so that each spatial position (h, w) yields a temporal sequence of length T with feature dimension d. Standard multi-head self-attention [52] is then applied to each of the $H' \times W'$ sequences independently:

$$\mathbf{F}^M_{0:T}(h, w) \leftarrow \mathrm{MHA}\big(\mathbf{F}^M_{0:T}(h, w)\big), \quad \forall\,(h, w), \tag{4}$$

where MHA denotes multi-head self-attention with query, key, and value all equal to the input. This design enables each spatial position to attend to its corresponding features across all T frames, capturing temporal dynamics such as storm movement and intensity evolution without mixing information across spatial locations. The output, combined with a residual connection, serves as the updated primary features for the next MTAI iteration, while the auxiliary features remain fixed.

Post-Fusion Encoder. After two MTAI iterations, the fused features $\mathbf{F}^M_{0:T} \in \mathbb{R}^{T\times d\times H'\times W'}$ are further processed by a post-fusion encoder consisting of a 3D convolution with spatial stride 2 (producing further $2\times$ spatial downsampling), two 3D residual blocks, and a $1\times1\times1$ projection, yielding the final cross-modal condition:

$$\mathbf{F}^C_{0:T} \in \mathbb{R}^{T\times d_{\text{out}}\times \frac{H_c}{4}\times \frac{W_c}{4}},$$

where d_{out} is the output channel dimension. This output serves as the conditioning input for the Satellite-to-Radar Diffusion Model.

C. Satellite-to-Radar Diffusion Model

To enhance the fidelity and detail of satellite-to-radar translation, we leverage the probabilistic framework of diffusion models [53], [54], [55]. As illustrated in Fig. 2, the SRDiff pipeline consists of two main stages: training and inference. During training, the ground truth radar sequence $x_{0:T}$ is first encoded into a latent feature space using a frame-wise Variational Autoencoder (VAE). Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ is then added to the latent feature $\mathbf{z}^x_{0:T}$, and the noisy latent feature is concatenated with the conditioned representation $\mathbf{F}^C_{0:T}$. The combined representation is then processed by the DiT model, which learns to predict and remove the noise component. During inference, the denoising process follows a reverse diffusion trajectory, starting from pure Gaussian noise $\mathbf{z}^x_N \sim \mathcal{N}(0, \mathbf{I})$ and iteratively refining it using the trained DiT model. The process follows a scheduled step size until reaching the predefined timestep N, progressively recovering the latent feature. Finally, the VAE decoder reconstructs the radar sequence $\hat{x}_{0:T}$ from the estimated clean latent representation. Apart from the VAE encoder-decoder, the entire pipeline, including DiT and the conditioning module, is trained end-to-end to ensure seamless integration and optimal performance.

Frame-wise Variational Autoencoder. To enhance computational efficiency while maintaining high-fidelity radar generation, we adopt a frame-wise VAE similar to [56]. The VAE compresses each radar frame into a low-dimensional latent, reducing the burden on the diffusion model. The radar sequence $x_{0:T}$ is encoded frame by frame into the latent space via encoder $\mathcal{E}$: $\mathbf{z}^{x_{0:T}} = \mathcal{E}(x_{0:T})$, where $\mathbf{z}^{x_{0:T}} \in \mathbb{R}^{T\times d\times H'\times W'}$ is the encoded representation. During denoising, the DiT model predicts the noise component: $v_\theta(\mathbf{z}^{x_{0:T}}, \mathbf{c}, t)$, where $\mathbf{c} = \mathbf{F}^C_{0:T}$ is the conditioning feature from CMCA, and θ denotes the learnable parameters of DiT. After iterative denoising, the latent is decoded by the VAE decoder $\mathcal{D}$ to reconstruct the final radar sequence: $\hat{x}_{0:T} = \mathcal{D}(\mathbf{z}^{x_{0:T}})$.

Training and Inference. Inspired by recent advances in diffusion models [57], [58], [59], we adopt rectified flow [60] over the conventional IDDPM sampling [56]. Rectified flow improves training stability and significantly reduces inference steps.

Instead of a standard noise-to-data mapping, rectified flow defines a linear transformation between the normal distribution and the data distribution in the latent space:

$$\mathbf{z}^x_t = (1 - t)\mathbf{z}^x_0 + t\mathbf{z}^x_1, \tag{5}$$

where $\mathbf{z}^x_1$ is a latent-space data sample, and $\mathbf{z}^x_0$ is a sample from a normal distribution.

The training objective minimizes the difference between the predicted velocity field and the true displacement:

$$\mathcal{L}(\theta) := \mathbb{E}_{\mathbf{z}^x_1, \mathbf{z}^x_0}\left[\|v_\theta(\mathbf{z}^x_t, t, \mathbf{c}) - (\mathbf{z}^x_1 - \mathbf{z}^x_0)\|^2_2\right]. \tag{6}$$

During inference, the sampling process is performed iteratively using a discretized rectified flow:

$$\mathbf{z}^x_{t-\frac{1}{N}} = \mathbf{z}^x_t - \frac{1}{N} v_\theta(\mathbf{z}^x_t, t, \mathbf{c}), \quad \forall t \in \{1, 2, \ldots, N\}/N. \tag{7}$$

This iterative denoising process refines the noisy latent feature until convergence, effectively reconstructing radar sequences with improved accuracy and temporal consistency.

D. Integration with Nowcasting Models

SRDiff is highly flexible and generalizable, offering seamless integration with existing precipitation nowcasting models to establish a complete satellite-to-radar forecasting pipeline. As shown in Fig. 3, SRDiff can be incorporated as a preprocessing module into radar-to-radar forecasting models. Specifically, SRDiff first generates radar-like inputs from satellite observations, which are then fed into a Radar Prediction Model for future radar forecasting. This integration extends radar-based nowcasting pipelines to regions without direct radar observations, enriching their input data and improving forecast reliability. By bridging the gap between satellite and radar data, SRDiff enhances the applicability and performance of existing nowcasting models, making short-term precipitation forecasting more accessible and effective across diverse environments.

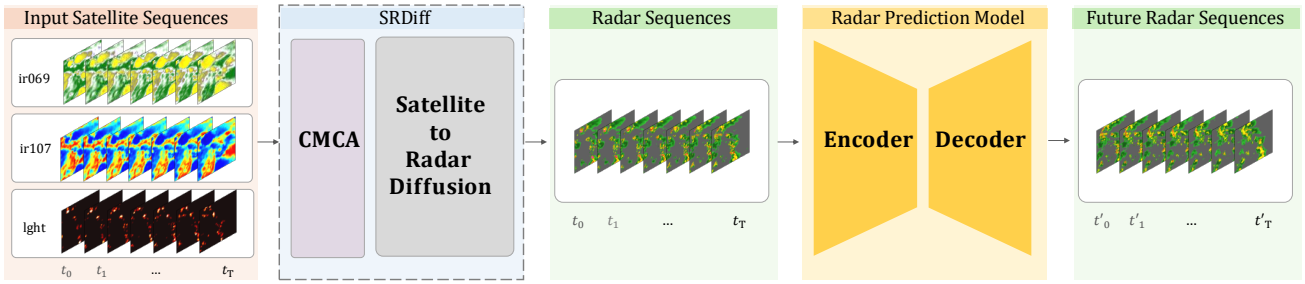

Fig. 3. Integration of SRDiff with Nowcasting Models. SRDiff serves as a preprocessing module by converting satellite observations into radar-like representations, which are then fed into a Radar Prediction Model for future radar forecasting. This integration extends radar-based nowcasting to regions without direct radar coverage, enhancing the accuracy and applicability of precipitation forecasting.

IV. Experiments

A. Datasets

We evaluate SRDiff on two public radar benchmark datasets, namely **SEVIR** [35] and **Sat2Rdr** [28]. **SEVIR** combines NEXRAD radar and GOES-16 satellite observations. The radar images in the SEVIR dataset are stored as integer values in the range of 0–254, with the value 255 indicating regions where data are missing. This encoded representation is generally well suited for deep learning-based methods, where normalization is typically applied to facilitate stable training and improve model performance. To recover the corresponding true units of Vertically Integrated Liquid (VIL, kg/m^2), the following decoding equations are applied.

$$R(x) = \begin{cases} 0, & x \leq 5, \\ \dfrac{x-2}{90.66}, & 5 < x \leq 18, \\ \exp\left(\dfrac{x-83.9}{38.9}\right), & 18 < x \leq 254. \end{cases}$$

VIL quantifies the total liquid water mass per unit area integrated through the atmospheric column, serving as a proxy for storm intensity and hail potential. Following Earthformer [5], we evaluate at encoded thresholds $x \in \{16, 74, 133, 160, 181, 219\}$, which correspond to the true VIL values $R(x)$ of approximately $\{0.15, 0.78, 3.53, 7.07, 12.13, 32.22\}$ kg/m^2, respectively [35].

SEVIR also provides aligned GOES-16 modalities: *ir069* (Water Vapor, 6.9 μm), *ir107* (Infrared, 10.7 μm), and *lght* (GLM lightning events, effective $\sim$8 km). For the satellite-to-radar setting, we follow the standard SEVIR temporal split with June 1, 2019 as the train-test cutoff: events from January 1, 2018 to May 31, 2019 are used for training, while events from June 1, 2019 onward are used for testing. We retain only events with complete *ir069*, *ir107*, *lght*, and *vil* observations and zero missing values, since some modalities are unavailable for certain SEVIR events. This yields 8,518 training events and 3,143 test events. Each 49-frame event is converted into three 25-frame radar-sequence samples using a stride of 12, resulting in 25,554 training samples and 9,429 test samples. Thus, samples denote sequences rather than individual image frames. For validation, we reuse the training split rather than defining an independent validation set.

TABLE II
Implementation Details and Training Hyper-Parameters

Parameter Name	Value
Input context length (VIL)	7 frames (70 min)
Forecast horizon (VIL)	6 frames (60 min)
Spatial downsampling	3×
Temporal downsampling	2× (ratio 0.5)
VAE input resolution	128 × 128
Optimizer for VAE	AdamW (lr=1×10^{-4}, cosine)
Diffusion noise schedule	Linear, 1000 steps
Inference denoising steps	10 (Rectified Flow)
Optimizer for diffusion	AdamW (lr=5×10^{-5}, cosine)
Training steps	200k
Hardware setup	4 × NVIDIA A100 GPUs
Global batch size	32
Sat2Rdr supervision	Single-frame translation

Sat2Rdr contains paired radar and satellite data, along with a 2-km Digital Elevation Model (DEM), covering September 2019–June 2024. The satellite channels are Infrared 10.5 μm and Water Vapor 6.3/7.3 μm, sampled hourly over a 1000 km $\times$ 1000 km domain at 2-km resolution. The radar product is Hybrid Surface Rainfall (HSR, mm/hr), a precipitation intensity field derived from merged Korean ground radar observations [28]. We use 2019–2023 for training and 2023–2024 for testing to ensure seasonal diversity. Sat2Rdr uses GK2A (GEO-KOMPSAT-2A), a geostationary satellite operated by the Korea Meteorological Administration (KMA), which is distinct from GOES-16. The channel differences between SEVIR and Sat2Rdr arise from the distinct spectral band configurations of their respective satellite platforms. Furthermore, since Sat2Rdr does not provide radar geolocation information, it is not feasible for users to retrieve corresponding spectral bands from alternative satellite platforms to augment the dataset.

B. Evaluation Metrics

We adopt multiple metrics to capture both spatial fidelity and probabilistic forecast quality, following prior work [29], [41]. Specifically, we report the Continuous Ranked Probability Score (CRPS), the Structural Similarity Index Measure (SSIM), the Heidke Skill Score (HSS), and the Critical Success Index

TABLE III
QUANTITATIVE COMPARISON OF SATELLITE-TO-RADAR TRANSLATION METHODS ON THE SEVIR DATASET. **BOLD**: BEST AND UNDERLINE: SECOND-BEST.

Model Type	Method	CRPS ↓	SSIM↑	HSS ↑	FVD ↓	CSI-M POOL ↑			CSI-181 (12.1 kg/m^2) ↑			CSI-219 (32.2 kg/m^2) ↑		
						1 × 1	2 × 2	4 × 4	1 × 1	2 × 2	4 × 4	1 × 1	2 × 2	4 × 4
UNet	Sat2Rad	0.0432	0.4618	0.4149	172.5	0.2980	0.3180	0.3401	0.2205	0.2466	0.2668	0.1052	0.1541	0.2130
GAN	CycleGAN	0.0414	0.5244	0.4435	136.7	0.3175	0.3492	0.3994	0.2569	0.3009	0.3646	0.1329	0.1789	0.2503
	Pix2Pix	0.0441	0.5552	0.4789	118.4	0.3487	0.3795	0.4260	0.2932	0.3382	0.3972	0.1304	0.1763	0.2375
	StegoGAN	0.0402	0.5748	0.4837	82.4	0.3518	0.3856	0.4396	0.2846	0.3285	0.3868	0.1245	0.1672	0.2258
Diffusion	CasCast	0.0476	0.5356	0.4592	67.3	0.3345	0.3812	0.4327	0.2735	0.3212	0.3877	0.1145	0.1713	0.2421
	SRDiff	**0.0380**	**0.6025**	**0.4971**	**48.2**	**0.3632**	**0.3977**	**0.4526**	**0.2971**	**0.3466**	**0.4188**	**0.1549**	**0.2105**	**0.2886**

(CSI) with severe-weather variants. HSS evaluates agreement between predictions and observations while accounting for both correct forecasts and errors, and is defined as

$$\mathrm{HSS} = \frac{2(TP \cdot TN - FN \cdot FP)}{(TP+FN)(FN+TN)+(TP+FP)(FP+TN)},$$

where TP, TN, FP, and FN denote the counts at a chosen threshold. CSI measures the fraction of correctly predicted events among all observed or forecasted events, and is given by

$$\mathrm{CSI} = \frac{TP}{TP+FN+FP}.$$

For SEVIR, we compute CSI at encoded VIL thresholds $\{16, 74, 133, 160, 181, 219\}$, which correspond to the true VIL units of $\{0.15, 0.78, 3.53, 7.07, 12.13, 32.22\}$ kg/m^2. We further report pooled CSI with 2×2 and 4×4 spatial pooling to enhance robustness, including CSI-M (mean across all thresholds), as well as CSI-181 and CSI-219. For Sat2Rdr, we follow the approach of Park *et al.* [28] and report CSI at rainfall thresholds of 1 mm, 4 mm, and 8 mm, which are operationally relevant for moderate-to-heavy rainfall and potential flood-related events. While informative, they are not specifically designed for satellite-to-radar translation and should therefore be interpreted with caution.

Implementation details. For the SEVIR dataset, following standard practice [39], [40], we predict 60 minutes of synthetic radar (7 frames) from satellite imagery and forecast the next 60 minutes of radar sequence (6 frames), with spatial and temporal resolutions downsampled by 3× and 2×, respectively. The VAE is trained on 128×128 (with a temporal downsampling ratio of 0.5) radar-derived VIL using AdamW (lr=1e-4, cosine schedule), similar to existing methods [29], [53]. For diffusion, we use a linear noise schedule (1000 steps), 10 denoising steps during inference with Rectified Flow [60], and train with AdamW (lr=5e-5, cosine schedule) for 200k steps on 4×NVIDIA A100 GPUs (global batch size 32). For the Sat2Rdr dataset, we adopt the same VAE and diffusion configurations but apply them to single-frame translation tasks. Since Sat2Rdr lacks temporal supervision, each HSR field is predicted independently based on the corresponding satellite input. The optimization-related parameters are summarized in Table II.

TABLE IV
QUANTITATIVE COMPARISON OF SATELLITE-TO-RADAR TRANSLATION METHODS ON SAT2RDR DATASET. **BOLD**: BEST AND UNDERLINE: SECOND-BEST.

Methods	CSI@1 mm	CSI@4 mm	CSI@8 mm
Synrad	0.1621	0.0305	0.0120
CycleGAN	0.2015	0.0514	0.0277
Pix2Pix	0.2243	0.0803	0.0589
StegoGAN	**0.2497**	0.1007	0.0601
BBDM	0.2121	0.0766	0.0511
SRDiff	0.2478	**0.1014**	**0.0631**

C. Comparison with State-of-the-Art.

To evaluate SRDiff's effectiveness in satellite-to-radar sequence generation, we compare it with five state-of-the-art models from different generative paradigms, namely Sat2Rad [26] (UNet-based), Pix2Pix [23], CycleGAN [24], and StegoGAN [25] (GAN-based), and CasCast [29] (Diffusion-based), on the SEVIR dataset. As shown in Table III, SRDiff consistently achieves the best performance across all evaluated metrics in terms of CRPS, SSIM, HSS, FVD, CSI-M, CSI-181, and CSI-219, demonstrating strong capability in both overall spatial fidelity and extreme weather capture. StegoGAN ranks second on several per-frame metrics (CRPS, SSIM, CSI-M), benefiting from its adversarial training for per-frame translation quality. However, it falls notably behind on high-intensity thresholds (CSI-181, CSI-219) and temporal coherence (FVD), as it lacks explicit temporal modeling and tends to under-represent rare extreme precipitation structures. A similar trend is observed for image-based translation models (CycleGAN and Pix2Pix), which, being frame-wise approaches, produce relatively low FVD scores. Although CasCast incorporates temporal modeling as a video-based method, it remains less competitive than our approach as it fails to incorporate explicit cross-modal attention to capture inter-modality correlations.

To assess generalization, we further evaluate on the Sat2Rdr dataset in Table IV, comparing SRDiff with Synrad, CycleGAN, Pix2Pix, StegoGAN [25], and BBDM [61]. SRDiff achieves the best CSI scores at 4 mm and 8 mm rainfall thresholds, and performs competitively at 1 mm, slightly trailing StegoGAN. These results highlight SRDiff's robustness in high-impact

TABLE V
PERFORMANCE COMPARISON WHEN INTEGRATED WITH DIFFERENT DOWNSTREAM NOWCASTING MODELS ON SEVIR DATASET. **BOLD**: BEST AND UNDERLINE: SECOND-BEST.

Nowcast Models	Method	CRPS ↓	SSIM ↑	HSS ↑	CSI-M POOL ↑			CSI-181 (12.1 kg/m^2) ↑			CSI-219 (32.2 kg/m^2) ↑		
					1 × 1	2 × 2	4 × 4	1 × 1	2 × 2	4 × 4	1 × 1	2 × 2	4 × 4
ConvLSTM	Sat2Rad	0.0588	0.3714	0.3080	0.2248	0.2273	0.2317	0.1024	0.0983	0.0948	0.0316	0.0374	0.0470
	CycleGAN	0.0540	0.4805	0.3568	0.2567	0.2673	0.2818	0.1530	0.1624	0.1735	**0.0630**	0.0779	0.1006
	Pix2Pix	0.0516	0.5251	0.3673	0.2664	0.2790	0.2953	**0.1546**	0.1656	0.1781	0.0505	0.0637	0.0848
	StegoGAN	0.0498	0.5326	0.3706	0.2698	0.2808	0.2976	0.1486	0.1648	0.1808	0.0518	0.0658	0.0872
	CasCast	0.0450	0.5605	0.3752	0.2724	0.2827	0.3005	0.1477	0.1640	0.1877	0.0543	0.0695	0.0963
	SRDiff	**0.0444**	**0.5650**	**0.3830**	**0.2783**	**0.2890**	**0.3073**	0.1510	**0.1660**	**0.1879**	0.0617	**0.0780**	**0.1056**
SimVP	Sat2Rad	0.0564	0.4644	0.3079	0.2235	0.2252	0.2288	0.1057	0.0998	0.0944	0.0370	0.0410	0.0472
	CycleGAN	0.0534	0.5100	0.3575	0.2561	0.2653	0.2791	0.1529	0.1606	0.1709	**0.0704**	**0.0822**	0.0990
	Pix2Pix	0.0510	0.5424	0.3638	0.2629	0.2722	0.2852	0.1520	0.1567	0.1635	0.0506	0.0591	0.0706
	StegoGAN	0.0492	0.5486	0.3724	0.2708	0.2804	0.2958	0.1536	0.1586	0.1678	0.0528	0.0618	0.0742
	CasCast	0.0450	0.5784	0.3848	0.2793	0.2899	0.3081	0.1572	0.1722	0.1952	0.0557	0.0701	0.0953
	SRDiff	**0.0444**	**0.5846**	**0.3936**	**0.2860**	**0.2973**	**0.3164**	**0.1619**	**0.1750**	**0.1965**	0.0645	0.0803	**0.1065**
TAU	Sat2Rad	0.0564	0.4387	0.3165	0.2300	0.2348	0.2418	0.1130	0.1138	0.1159	0.0363	0.0433	0.0540
	CycleGAN	0.0522	0.5182	0.3637	0.2607	0.2740	0.2944	0.1558	0.1716	0.1938	**0.0728**	**0.0908**	**0.1184**
	Pix2Pix	0.0492	0.5486	0.3638	0.2599	0.2685	0.2809	0.1395	0.1456	0.1545	0.0377	0.0463	0.0580
	StegoGAN	0.0478	0.5536	0.3742	0.2696	0.2806	0.2958	0.1428	0.1498	0.1608	0.0426	0.0518	0.0648
	CasCast	0.0449	0.5815	0.3903	0.2790	0.2888	0.3061	0.1466	0.1581	0.1776	0.0509	0.0646	0.0869
	SRDiff	**0.0438**	**0.5857**	**0.3936**	**0.2832**	**0.2988**	**0.3250**	**0.1646**	**0.1885**	**0.2252**	0.0594	0.0797	0.1148

TABLE VI
DIRECT SATELLITE NOWCASTING ON SEVIR (4×4 POOLING). EACH $+x$ MIN COLUMN REPORTS THE RUNNING CUMULATIVE AVERAGE CSI-M OVER ALL PREDICTED FRAMES FROM $+10$ MIN UP TO $+x$ MIN.

Models	Cumulative CSI-M ↑					
	≤+10 min	≤+20 min	≤+30 min	≤+40 min	≤+50 min	≤+60 min
Sat2Rad	0.307	0.289	0.280	0.267	0.254	0.237
SimVP	0.358	0.333	0.311	0.294	0.279	0.266
TAU	0.382	0.363	0.345	0.330	0.316	0.305
SRDiff	**0.392**	**0.377**	**0.362**	**0.348**	**0.335**	**0.325**

scenarios, especially for localized heavy precipitation, and its strong generalization to radar-sparse regions.

D. Integration with Downstream Nowcasting

As detailed in the methods, we evaluate the utility of SRDiff as a satellite-to-radar translation module by plugging it into three representative nowcasting backbones, namely ConvLSTM [4], SimVP [34], and TAU [33]. Following [29], each backbone is trained on SEVIR using 7 historical frames to predict the next 6 frames, with the same train-test split as the translation setup. We then feed these trained models with synthetic radar sequences translated from satellite observations by Sat2Rad, Pix2Pix, CycleGAN, StegoGAN, CasCast, and SRDiff. As shown in Table V, integrating SRDiff yields consistent downstream gains across nearly all metrics. In particular, SRDiff attains the best CRPS, SSIM, HSS, and CSI-M, reflecting stronger spatiotemporal coherence and improved structural fidelity. Notably, StegoGAN outperforms Pix2Pix on most downstream metrics (CRPS, SSIM, HSS, CSI-M), reflecting its stronger per-frame translation quality. However, StegoGAN still trails CasCast and SRDiff by a clear margin, particularly on CSI-181 and CSI-219 at larger pooling scales, confirming that temporal coherence, in addition to per-frame fidelity, is a critical factor for downstream nowcasting utility. In contrast, SRDiff's explicit temporal modeling produces smooth, coherent input sequences that downstream models can effectively exploit, yielding consistent gains across all three nowcasting backbones.

To further assess the viability of direct satellite-based nowcasting, we compare our two-stage approach against end-to-end models that predict future radar fields directly from satellite inputs under the 4×4 pooling setting. Table VI reports the *running cumulative average* of CSI-M: each $\leq+x$ min column gives the mean CSI-M computed over all predicted frames from $+10$ min up to $+x$ min, providing a progressively inclusive measure of forecast quality over an expanding evaluation window. As the horizon extends, later frames, which are inherently harder to predict, gradually lower the cumulative average. SRDiff achieves the highest cumulative CSI-M at every horizon, consistently surpassing SimVP, TAU, and the single-stage Sat2Rad baseline. Notably, the performance margin widens at longer horizons ($\leq+50$ and $\leq+60$ min), indicating that the temporally coherent radar sequences produced by SRDiff compound their benefit as the forecast window grows.

E. Visualization.

Beyond quantitative scores, visual consistency and physical interpretability are critical in satellite-to-radar translation. Fig. 4 compares satellite inputs (ir069, ir107, lght), ground-truth

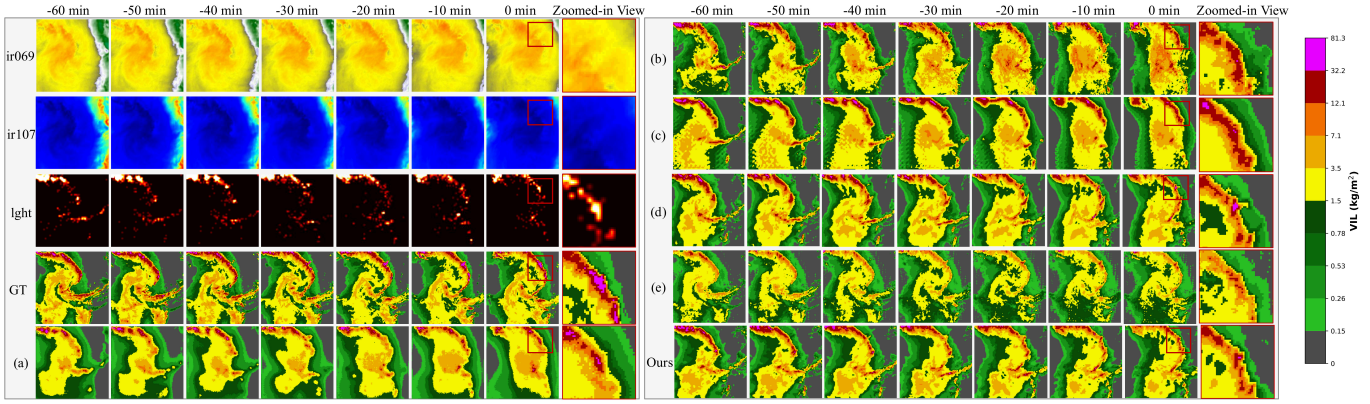

Fig. 4. Comparison of satellite-to-radar translation results. Left: satellite inputs (ir069, ir107, lght), ground-truth radar-derived VIL field, and (a) Sat2Rad. Right: outputs from (b) CycleGAN, (c) Pix2Pix, (d) CasCast, (e) StegoGAN, and ours (top to bottom). Rightmost column shows zoom-in at 0 min for visual comparison. Colorbars indicate VIL intensity in kg/m^2.

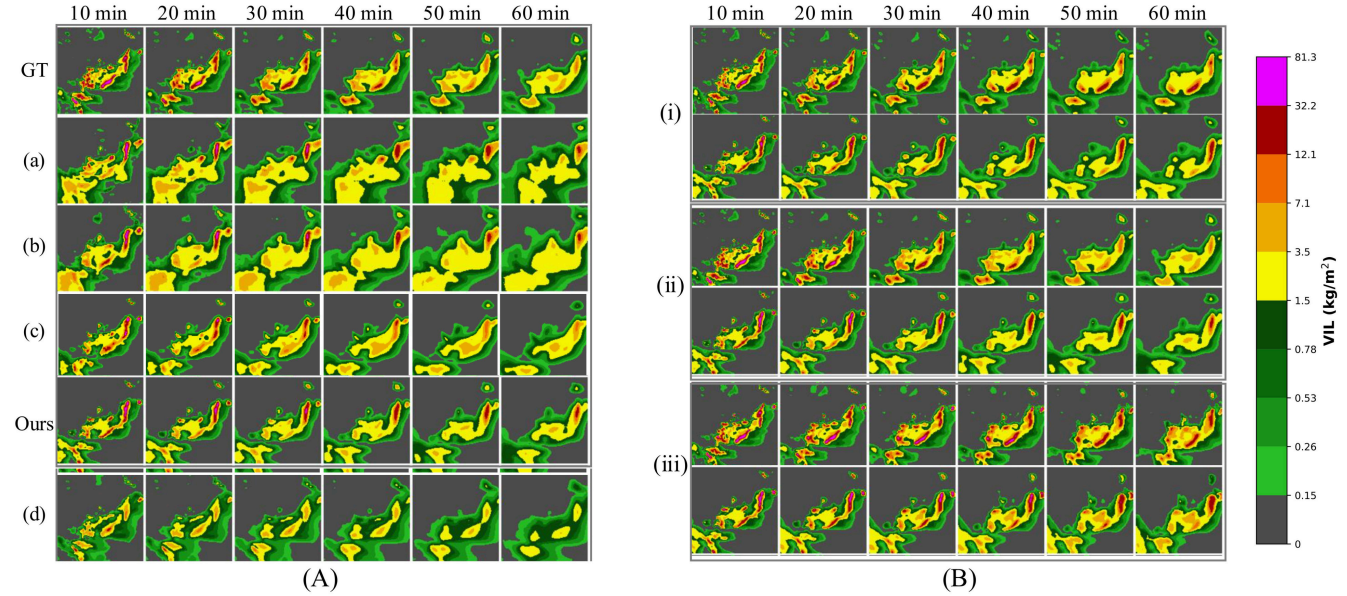

Fig. 5. (A) Qualitative comparison of prediction results when integrated with SimVP model. From top to bottom: Ground Truth, (a) CycleGAN, (b) Pix2Pix, (c) CasCast, (d) StegoGAN, and our method. (B) Qualitative comparison of prediction results obtained by our method when integrated with (i) ConvLSTM, (ii) SimVP, and (iii) TAU. For each method, the top row shows the ground truth, and the bottom row shows the results from our method. Colorbars indicate VIL intensity in kg/m^2.

radar, and outputs from Sat2Rad, CycleGAN, Pix2Pix, CasCast, StegoGAN, and SRDiff. SRDiff yields sharper structures and better continuity, especially in intense cores. Fig. 5 further illustrates the benefits of integrating SRDiff into downstream radar forecasting models. Fig. 5 (A) compares results when paired with SimVP, where SRDiff consistently delivers sharper, more coherent predictions in shape and intensity. These improvements align with our core hypothesis discussed in the previous section, confirming SRDiff's ability to narrow the modality gap and enhance radar prediction quality. Fig. 5 (B) demonstrates the compatibility of SRDiff with different downstream models, where the overall forecasting performance varies depending on the strength of the underlying forecaster, highlighting SRDiff's robustness and its potential as a general-purpose satellite-to-radar enhancement module.

To illustrate behavior on a concrete event, Fig. 6 shows a SEVIR case where, relative to GAN and UNet baselines, SRDiff preserves compact convective cores, maintains sharp reflectivity gradients, and suppresses blocky/grid artifacts, yielding temporally consistent sequences with less frame-to-frame flicker and centroid jitter. As a downstream counterpart, Fig. 7 presents SimVP-based radar-to-radar forecasts for the same event, where inputs translated by SRDiff retain elongated storm geometry and rainband continuity over longer lead times with better preservation of high-intensity gradients. These qualitative patterns align with the quantitative gains in Tables III and V such as the strong performance at CSI-181 and CSI-219 across different pooling resolutions, especially at +50 and +60 min. The ablations in Table VII attribute the improvements to narrowing the satellite–radar modality gap upstream, which in turn produces more coherent and skillful downstream forecasts under severe weather.

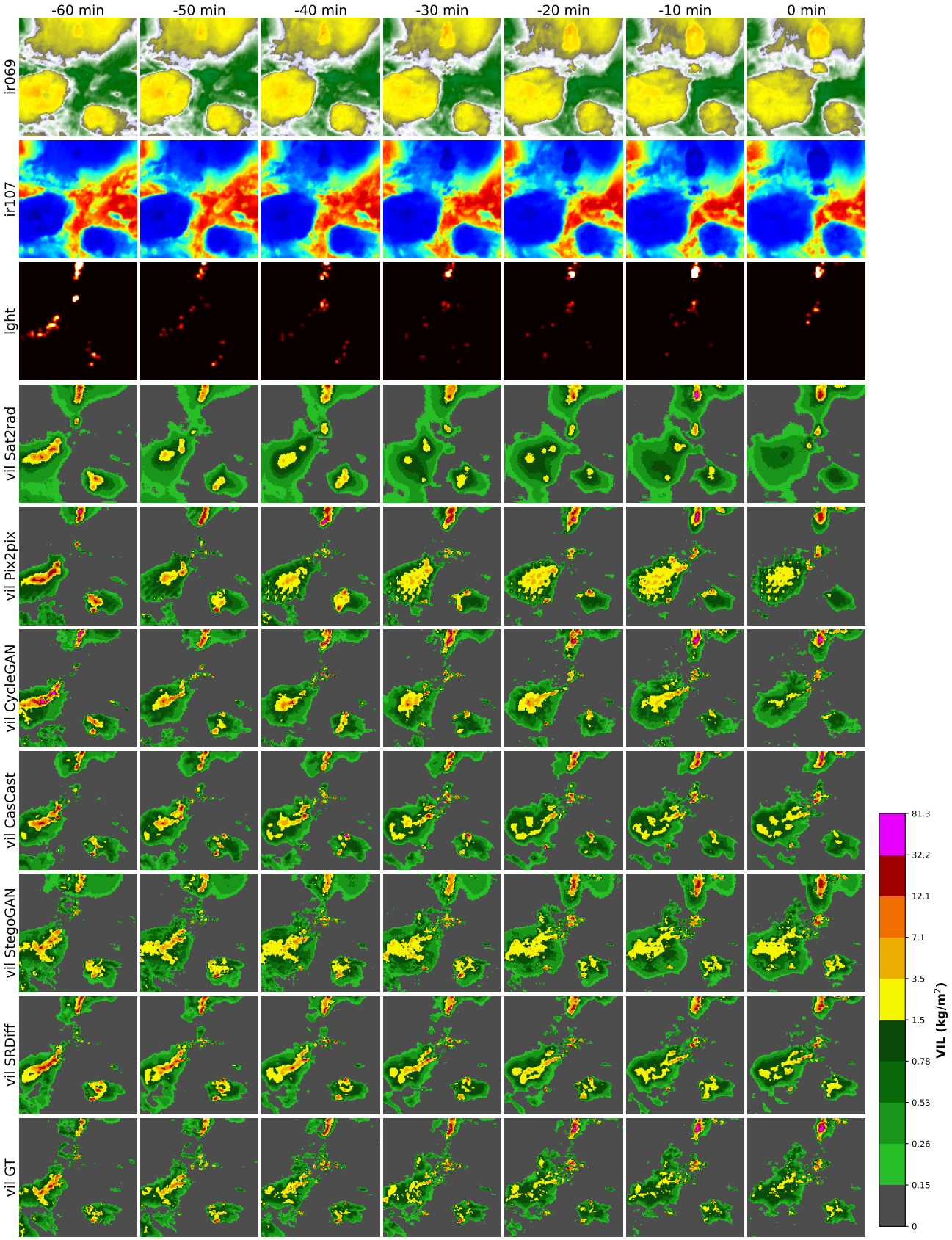

Fig. 6. Satellite-to-radar translation on SEVIR event. Columns depict consecutive time steps; rows show *Ground Truth* and outputs from *Sat2Rad*, *CycleGAN*, *Pix2Pix*, *CasCast*, *StegoGAN*, and **SRDiff**. SRDiff preserves compact convective cores and filamentary rainband structures while mitigating blocky/grid artifacts and temporal flicker observed in GAN/UNet baselines. Colorbars indicate VIL intensity in kg/m^2.

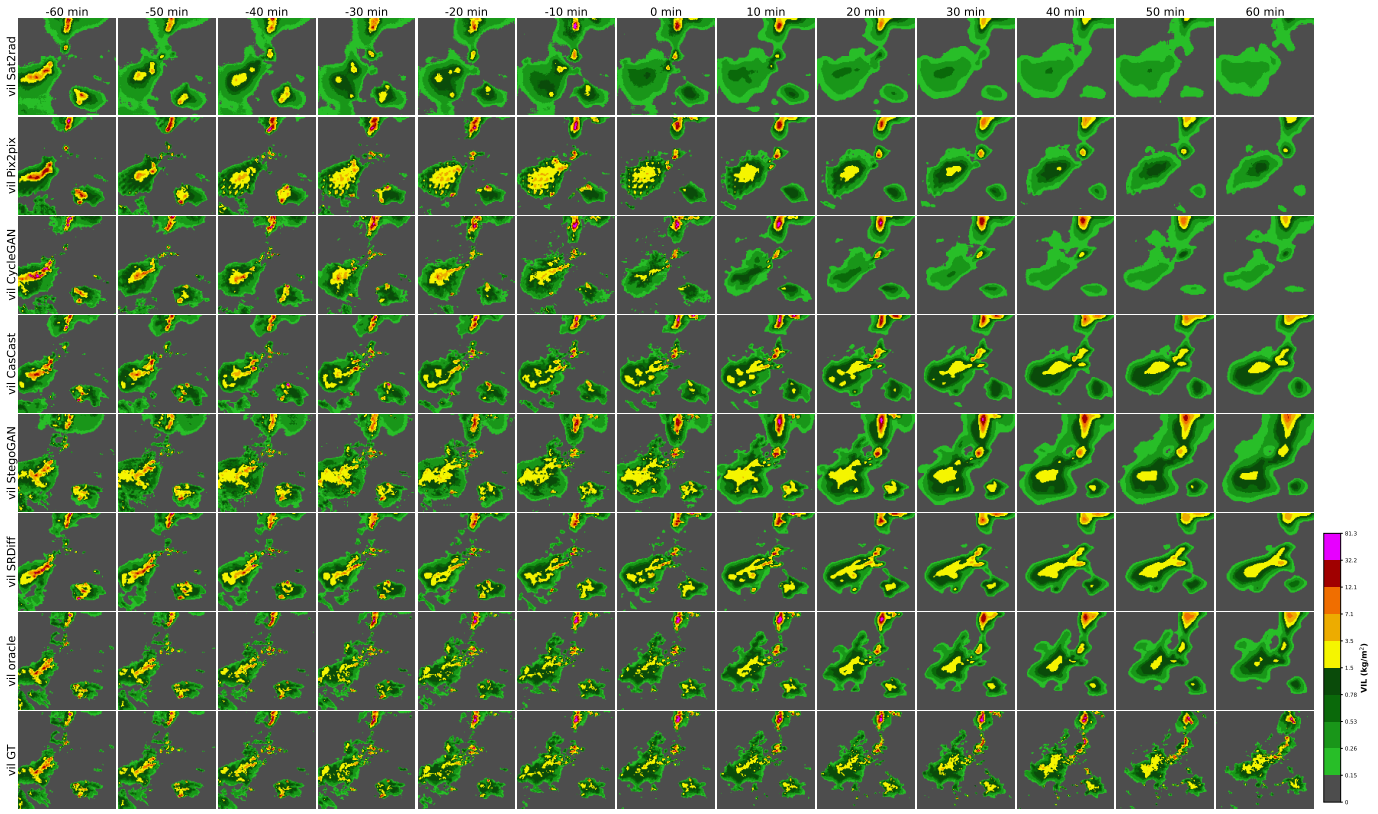

Fig. 7. **Downstream radar-to-radar nowcasting on a SEVIR event (SimVP).** Feeding SimVP with radar sequences translated from satellite by Sat2Rad, Pix2Pix, CycleGAN, CasCast, StegoGAN, and SRDiff, the forecasts driven by *Oracle* denotes SimVP run on *ground-truth radar inputs* (same training, identical setup), serving as an upper-bound reference rather than an operational baseline. This case study complements the qualitative summary in Fig. 5. Colorbars indicate VIL intensity in kg/m^2.

TABLE VII
ABLATION RESULTS ON SEVIR DATASET. CMCA TARGETS THE MODALITY GAP, WHILE MAM ADDRESSES TEMPORAL DEPENDENCY.

MG	TG	CRPS ↓	SSIM ↑	HSS ↑	CSI-K ↑						
					16	74	133	160	181	219	avg
		0.0434	0.5658	0.4694	0.6011	0.4664	0.2878	0.2878	0.2733	0.1300	0.3411
✓		0.0399	0.5859	0.4752	0.6139	0.4737	0.2876	0.2921	0.2726	0.1318	0.3453
	✓	0.0385	0.5883	0.4869	0.6189	0.4831	0.2933	0.3052	0.2882	0.1425	0.3552
✓	✓	**0.0380**	**0.6025**	**0.4971**	**0.6244**	**0.4872**	**0.3036**	**0.3135**	**0.2971**	**0.1549**	**0.3632**

F. Ablation Study on Modality Gap and Temporal Dependency.

To tackle the key challenges of modality gap and temporal dependency, we design targeted components within our framework. To evaluate their individual contributions, we conduct ablation experiments, with results summarized in Table VII. Removing either component results in a significant performance degradation, confirming that both are essential for achieving optimal results. Only when both modules are present does the model reach its full potential, highlighting their importance in effectively bridging the modality gap and modeling temporal dynamics.

To further quantify temporal coherence, we report Fréchet Video Distance (FVD) [62] for both baseline methods and ablation variants of our model in Table VIII. The full SRDiff model achieves the lowest FVD (48.2), followed by CasCast (67.3),

TABLE VIII
TEMPORAL COHERENCE COMPARISON (FVD [62]). LOWER FVD INDICATES BETTER TEMPORAL CONSISTENCY ACROSS GENERATED RADAR SEQUENCES.

Model / Variant	**FVD ↓**	**Δ vs. SRDiff**
Baseline Methods		
Sat2Rad	172.5	+257.9%
CycleGAN	136.7	+183.6%
Pix2Pix	118.4	+145.6%
StegoGAN	82.4	+71.0%
CasCast	67.3	+39.6%
SRDiff (Full)	**48.2**	—
Ablation Variants		
w/o Temporal Attention	237.6	+393.0%
w/o Cross-Modal Adapter	72.6	+50.6%

TABLE IX
BIDIRECTIONAL CROSS-REGION GENERALIZATION ON SEVIR. TRAINING ON ONE REGION AND TESTING ON THE OTHER DEMONSTRATES SRDIFF'S ROBUSTNESS ACROSS METEOROLOGICAL REGIMES. **BOLD**: BEST, UNDERLINE: SECOND-BEST PER SETTING.

Train	Test	Method	CSI-M ↑	CSI-16 ↑	CSI-74 ↑	CSI-219 ↑	SSIM ↑
Northern US	Northern US (In-Dom.)	Pix2Pix	0.3168	0.5452	0.4128	0.1162	0.5178
		StegoGAN	0.3426	0.5742	0.4485	0.1108	0.5486
		SRDiff	**0.3513**	**0.5916**	**0.4642**	**0.1359**	**0.5621**
	Southern US	Pix2Pix	0.2915	0.5489	0.3772	0.1213	0.4907
		StegoGAN	0.3098	0.5682	0.4016	0.0982	0.5136
		SRDiff	**0.3247**	**0.5973**	**0.4234**	**0.1435**	**0.5324**
Southern US	Southern US (In-Dom.)	Pix2Pix	0.3195	0.5804	0.4225	0.1308	0.5174
		StegoGAN	0.3385	0.6108	0.4498	0.1064	0.5348
		SRDiff	**0.3543**	**0.6302**	**0.4740**	**0.1545**	**0.5619**
	Northern US	Pix2Pix	0.3142	0.5418	0.4195	0.1071	0.5130
		StegoGAN	0.3356	0.5724	0.4416	0.1028	0.5348
		SRDiff	**0.3497**	**0.5883**	**0.4707**	**0.1265**	**0.5572**

TABLE X
CROSS-REGION DATASET COMPOSITION. GEOGRAPHIC SPLIT OF SEVIR EVENTS AT 38°N LATITUDE.

Region	Latitude Range	Train (< 2019-06)		Test (≥ 2019-06)	
		Events	Samples	Events	Samples
Northern US (≥ 38°N)	38.00–49.69°N	3,955	11,865	1,568	4,704
Southern US (< 38°N)	23.46–38.00°N	4,563	13,689	1,575	4,725
Total	23.46–49.69°N	**8,518**	**25,554**	**3,143**	**9,429**

while StegoGAN (82.4) ranks third, which is notably worse than both temporally-aware methods despite its strong per-frame generation quality. This confirms that frame-independent translation, regardless of single-frame fidelity, introduces temporal discontinuities that degrade sequence-level coherence. As image-based approaches, Sat2Rad, CycleGAN, Pix2Pix, and StegoGAN lack mechanisms to enforce inter-frame consistency, resulting in FVD substantially higher than CasCast and SRDiff. The "w/o Temporal Attention" variant exhibits an FVD of 237.6 (a 393% increase), demonstrating that temporal modeling provides consistent rather than occasional improvements. We further computed FVD across 1,000 randomly sampled test sequences, across which the full model achieved lower FVD in 94.3% of cases, confirming the robustness of the temporal attention mechanism.

G. Cross-Region Generalization

To evaluate robustness beyond in-distribution settings, we conduct bidirectional cross-region experiments within the SEVIR dataset. We split the SEVIR catalog by projection center latitude at 38°N, yielding two geographically distinct subsets: **Northern US** (≥38°N; Great Lakes, Upper Midwest, Northern Plains—continental convection and lake-effect precipitation) and **Southern US** (<38°N; Gulf Coast, Southeast, Southern Plains—tropical moisture influences and mesoscale convective systems). Only events with zero missing data across all four modalities (ir069, ir107, lght, vil) are included. We report the results in Table IX, with the corresponding statistics of the two subsets summarized in Table X.

We test generalization in both directions, *i.e.*, training on one region and evaluating on the other. As shown in Table IX, SRDiff consistently achieves the best performance across all metrics in every cross-region setting, including CSI-M, CSI-16, CSI-74, CSI-219, and SSIM. StegoGAN ranks second on CSI-M, CSI-16, CSI-74, and SSIM, indicating reasonable but not dominant per-frame translation quality. The cross-region CSI-M drops for SRDiff are modest: North→South decreases by 7.6% (0.3513→0.3247) and South→North by only 1.3% (0.3543→0.3497), averaging ∼4.5%, suggesting SRDiff learns region-agnostic cues from multi-modal satellite inputs rather than overfitting to local storm appearance. In contrast, StegoGAN exhibits a larger average performance drop in CSI-M (∼5.3%), indicating that SRDiff's temporal modeling also contributes to more stable generalization across meteorological regimes. The transfer asymmetry is meteorologically plausible: Southern US events more often involve tropical moisture transport and organized mesoscale convective systems (MCSs), whose spatial organization and convective–stratiform structure can differ from Northern continental convection (and lake-effect related precipitation), making North-trained models slightly less compatible when applied southward.

V. CONCLUSION AND FUTURE WORK

In this paper, we formalize satellite-to-radar translation as an operationally important yet underexplored problem and present SRDiff, a cross-modal, sequence-aware diffusion model tailored to this task. By aligning heterogeneous satellite channels with a Cross-Modal Conditional Adapter and modeling long-range dynamics via a Diffusion Transformer, SRDiff produces stable, physically plausible radar-like sequences. Extensive evaluations on SEVIR and Sat2Rdr show that SRDiff consistently surpasses deterministic and diffusion baselines across

multiple metrics. When integrated into nowcasting models such as ConvLSTM, SimVP and TAU, the converted radar fields deliver consistent gains, particularly on critical success indices (CSI-M, CSI-181, and CSI-219). These improvements extend the applicability of mature radar-centric models to radar-sparse regions and demonstrate clear operational value.

Limitations. Our current evaluation relies on two publicly available datasets (SEVIR and Sat2Rdr) with fixed channel configurations determined by their respective satellite platforms (GOES-16 and GK2A). In addition, Sat2Rdr does not provide geographic coordinates, preventing users from independently selecting or augmenting spectral bands. Consequently, cross-satellite transfer (*e.g.*, training on GOES-16 and testing on GK2A) remains challenging due to the differences in spectral responses, geographic coverage, derived radar products (VIL vs. HSR), and temporal resolutions. Although CMCA's channel-agnostic design mitigates this by requiring only modifications to the input projection layer, direct zero-shot cross-platform generalization has not been demonstrated and constitutes an important direction for future work.

In the future, we plan to incorporate deterministic priors, building an end-to-end forecasting pipeline, and exploring additional meteorological signals for enhanced spatiotemporal representation. We will further examine the model's generalization capability across diverse geographic regions and satellite platforms by conducting evaluations on tropical datasets (*e.g.*, Singapore) and by extending experiments to additional geostationary satellite systems such as Himawari. Beyond nowcasting, we will further investigate downstream applications through integration with NWP forecasts, enabling seamless temporal model blending for consistent forecasting across extended lead times.

VI. Acknowledgment

This research is supported by the National Research Foundation, Singapore and the National Environment Agency, Singapore under the Weather Science Research Programme (Award No. WSRP-2025-1R-03-03). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore and the National Environment Agency.

References

[1] E. Saltikoff, K. Friedrich, J. Soderholm, K. Lengfeld, B. Nelson, A. Becker, R. Hollmann, B. Urban, M. Heistermann, and C. Tassone, "An overview of using weather radar for climatological studies: Successes, challenges, and potential," *Bulletin of the American Meteorological Society*, vol. 100, no. 9, pp. 1739–1752, 2019.

[2] N. McCarthy, A. Guyot, A. Dowdy, and H. McGowan, "Wildfire and weather radar: A review," *Journal of Geophysical Research: Atmospheres*, vol. 124, no. 1, pp. 266–286, 2019.

[3] Z. Sokol, J. Szturc, J. Orellana-Alvear, J. Popova, A. Jurczyk, and R. Célleri, "The role of weather radar in rainfall estimation and its application in meteorological and hydrological modelling—a review," *Remote Sensing*, vol. 13, no. 3, p. 351, 2021.

[4] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[5] Z. Gao, X. Shi, H. Wang, Y. Zhu, Y. B. Wang, M. Li, and D.-Y. Yeung, "Earthformer: Exploring space-time transformers for earth system forecasting," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 390–25 403, 2022.

[6] X. Shi, Z. Gao, L. Lausen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, "Deep learning for precipitation nowcasting: A benchmark and a new model," *Advances in neural information processing systems*, vol. 30, 2017.

[7] Y. Wang, L. Jiang, M.-H. Yang, L.-J. Li, M. Long, and L. Fei-Fei, "Eidetic 3d lstm: A model for video prediction and beyond," in *International conference on learning representations*, 2018.

[8] C. Bai, F. Sun, J. Zhang, Y. Song, and S. Chen, "Rainformer: Features extraction balanced network for radar-based precipitation nowcasting," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[9] B. P. Shukla, C. M. Kishtawal, and P. K. Pal, "Prediction of satellite image sequence for weather nowcasting using cluster-based spatiotemporal regression," *IEEE transactions on geoscience and remote sensing*, vol. 52, no. 7, pp. 4155–4160, 2013.

[10] X. Dong, Z. Zhao, Y. Wang, J. Wang, and C. Hu, "Motion-guided global–local aggregation transformer network for precipitation nowcasting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[11] Y. Wang, L. Yao, H. Jiang, T. Liu, Y. Lu, and C. Zhou, "Precipitation nowcasting based on radar echo images via muti-scale spatiotemporal lstm," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[12] Y. Yin, S. Chen, Y. Li, L. Wang, R. Jin, W. Cui, and S. Xiang, "Simcast: Enhancing precipitation nowcasting with short-to-long term knowledge distillation," in *ICME*, 2025, pp. 1–6.

[13] C. Huang, P. Mu, J. Zhang, D. Xian, L. Dong, and C. Bai, "Benchmark dataset and deep learning method for global tropical cyclone forecasting," *Nature Communications*, vol. 16, p. 5923, 2025.

[14] C. Huang, P. Mu, C. Bai, and P. A. G. Watson, "TCP-diffusion: A multi-modal diffusion model for global tropical cyclone precipitation forecasting with change awareness," in *Proceedings of the 42nd International Conference on Machine Learning*, ser. PMLR, vol. 267, 2025, pp. 25 634–25 653.

[15] Federal Aviation Administration, "FAA's NextGen Weather Processor (NWP)," 2020, accessed: 2020-05-28. [Online]. Available: https://www.faa.gov/nextgen/programs/weather/nwp/

[16] G. J. Huffman, D. T. Bolvin, D. Braithwaite, K. Hsu, R. J. Joyce, C. Kidd, E. J. Nelkin, S. Sorooshian, E. F. Stocker, J. Tan *et al.*, "Integrated multi-satellite retrievals for the global precipitation measurement (gpm) mission (imerg)," *Satellite precipitation measurement: Volume 1*, pp. 343–353, 2020.

[17] R. J. Joyce, J. E. Janowiak, P. A. Arkin, and P. Xie, "Cmorph: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution," *Journal of hydrometeorology*, vol. 5, no. 3, pp. 487–503, 2004.

[18] S. Sorooshian, K.-L. Hsu, X. Gao, H. V. Gupta, B. Imam, and D. Braithwaite, "Evaluation of persiann system satellite-based estimates of tropical rainfall," *Bulletin of the American Meteorological Society*, vol. 81, no. 9, pp. 2035–2046, 2000.

[19] A. Y. Hou, R. K. Kakar, S. Neeck, A. A. Azarbarzin, C. D. Kummerow, M. Kojima, R. Oki, K. Nakamura, and T. Iguchi, "The global precipitation measurement mission," *Bulletin of the American Meteorological Society*, vol. 95, no. 5, pp. 701–722, 2014.

[20] T. J. Schmit, S. S. Lindstrom, J. J. Gerth, and M. M. Gunshor, "Applications of the 16 spectral bands on the advanced baseline imager (abi)," 2018.

[21] C. Luo, X. Li, Y. Ye, S. Feng, and M. K. Ng, "Experimental study on generative adversarial network for precipitation nowcasting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022.

[22] Z. Zhao, X. Dong, Y. Wang, and C. Hu, "Advancing realistic precipitation nowcasting with a spatiotemporal transformer-based denoising diffusion model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[24] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[25] S. Wu, Y. Chen, S. Mermet, L. Hurni, K. Schindler, N. Gonthier, and L. Landrieu, "Stegogan: Leveraging steganography for non-bijective

image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7922–7931.
[26] J. Pihrt, R. Raevskiy, P. Šimánek, and M. Choma, "Weatherfusionnet: Predicting precipitation from satellite data," *arXiv preprint arXiv:2211.16824*, 2022.
[27] X. Yu, Y. Li, J. Ma, C. Li, and H. Wu, "Diffusion-rscc: Diffusion probabilistic model for change captioning in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–13, 2025.
[28] Y.-J. Park, D. Kim, M. Seo, H.-G. Jeon, and Y. Choi, "Data-driven precipitation nowcasting using satellite imagery," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 27, 2025, pp. 28 284–28 292.
[29] J. Gong, L. Bai, P. Ye, W. Xu, N. Liu, J. Dai, X. Yang, and W. Ouyang, "Cascast: Skillful high-resolution precipitation nowcasting via cascaded modelling," in *Proceedings of the 41st International Conference on Machine Learning*, ser. PMLR, vol. 235, 2024, pp. 15 809–15 822.
[30] J. Pathak, S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli *et al.*, "Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators," *arXiv preprint arXiv:2202.11214*, 2022.
[31] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, "Accurate medium-range global weather forecasting with 3d neural networks," *Nature*, vol. 619, no. 7970, pp. 533–538, 2023.
[32] K. Chen, T. Han, J. Gong, L. Bai, F. Ling, J.-J. Luo, X. Chen, L. Ma, T. Zhang, R. Su *et al.*, "The operational medium-range deterministic weather forecasting can be extended beyond a 10-day lead time," *Communications Earth & Environment*, vol. 6, p. 518, 2025.
[33] C. Tan, Z. Gao, L. Wu, Y. Xu, J. Xia, S. Li, and S. Z. Li, "Temporal attention unit: Towards efficient spatiotemporal predictive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 770–18 782.
[34] Z. Gao, C. Tan, L. Wu, and S. Z. Li, "Simvp: Simpler yet better video prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3170–3180.
[35] M. Veillette, S. Samsi, and C. Mattioli, "Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 009–22 019, 2020.
[36] C. Kummerow, W. Barnes, T. Kozu, J. Shiue, and J. Simpson, "The tropical rainfall measuring mission (trmm) sensor package," *Journal of atmospheric and oceanic technology*, vol. 15, no. 3, pp. 809–817, 1998.
[37] Y. Wang, M. Long, J. Wang, Z. Gao, and P. S. Yu, "Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms," *Advances in neural information processing systems*, vol. 30, 2017.
[38] V. L. Guen and N. Thome, "Disentangling physical dynamics from unknown factors for unsupervised video prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 474–11 484.
[39] S. Ravuri, K. Lenc, M. Willson, D. Kangin, R. Lam, P. Mirowski, M. Fitzsimons, M. Athanassiadou, S. Kashem, S. Madge *et al.*, "Skilful precipitation nowcasting using deep generative models of radar," *Nature*, vol. 597, no. 7878, pp. 672–677, 2021.
[40] Y. Zhang, M. Long, K. Chen, L. Xing, R. Jin, M. I. Jordan, and J. Wang, "Skilful nowcasting of extreme precipitation with nowcastnet," *Nature*, vol. 619, no. 7970, pp. 526–532, 2023.
[41] Z. Gao, X. Shi, B. Han, H. Wang, X. Jin, D. Maddix, Y. Zhu, M. Li, and Y. B. Wang, "Prediff: Precipitation nowcasting with latent diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 78 621–78 656, 2023.
[42] D. Yu, X. Li, Y. Shen, J. Chen, X. Liu, and Q. Liu, "Diffcast: A unified framework via residual diffusion for precipitation nowcasting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 758–27 767.
[43] J. Wang, X. Wang, Y. Zhang, and J. Sun, "Skillful precipitation nowcasting using physical-driven diffusion networks," *Geophysical Research Letters*, vol. 51, no. 21, p. e2024GL110832, 2024.
[44] Y. Huang, P. Yang, F. Qin, C. Li, and X. Ling, "Rndiff: Rainfall nowcasting with condition diffusion model," *Pattern Recognition*, vol. 160, p. 111193, 2025.
[45] Z. Pan, R. Hang, Q. Liu, and X. Yuan, "A short-long term sequence learning network for precipitation nowcasting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.
[46] Ç. Küçük, A. Giannakos, S. Schneider, and A. Jann, "Transformer-based nowcasting of radar composites from satellite images for severe weather," *Artificial Intelligence for the Earth Systems*, vol. 3, no. 2, p. e230041, 2024.
[47] D. Torbunov, Y. Huang, H.-H. Tseng, H. Yu, J. Huang, S. Yoo, M. Lin, B. Viren, and Y. Ren, "Uvcgan v2: An improved cycle-consistent gan for unpaired image-to-image translation," 2023. [Online]. Available: https://arxiv.org/abs/2303.16280
[48] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–10.
[49] B. Li, K. Xue, B. Liu, and Y.-K. Lai, "Bbdm: Image-to-image translation with brownian bridge diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, 2023, pp. 1952–1961.
[50] Z. Luo, D. Chen, Y. Zhang, Y. Huang, L. Wang, Y. Shen, D. Zhao, J. Zhou, and T. Tan, "Videofusion: Decomposed diffusion models for high-quality video generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 209–10 218.
[51] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," in *ICLR*, 2017.
[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.
[53] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
[54] E. Hoogeboom, J. Heek, and T. Salimans, "simple diffusion: End-to-end diffusion for high resolution images," in *International Conference on Machine Learning*. PMLR, 2023, pp. 13 213–13 232.
[55] A. Gupta, L. Yu, K. Sohn, X. Gu, M. Hahn, F.-F. Li, I. Essa, L. Jiang, and J. Lezama, "Photorealistic video generation with diffusion models," in *European Conference on Computer Vision*. Springer, 2024, pp. 393–411.
[56] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*. PMLR, 2021, pp. 8162–8171.
[57] Z. Zheng, X. Peng, T. Yang, C. Shen, S. Li, H. Liu, Y. Zhou, T. Li, and Y. You, "Open-sora: Democratizing efficient video production for all," *arXiv preprint arXiv:2412.20404*, 2024.
[58] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.
[59] A. Helbling, T. H. S. Meral, B. Hoover, P. Yanardag, and D. H. Chau, "Conceptattention: Diffusion transformers learn highly interpretable features," in *Proceedings of the 42nd International Conference on Machine Learning*, ser. PMLR, vol. 267, 2025, pp. 22 946–22 963.
[60] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," in *International Conference on Learning Representations*, 2023.
[61] B. Li, K. Xue, B. Liu, and Y.-K. Lai, "Bbdm: Image-to-image translation with brownian bridge diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2205.07680
[62] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," in *International Conference on Learning Representations*, 2019.

You Qin received B.S. degree in Mathematics for Information and Computing Science at University of Electronic Science and Technology of China, Chengdu, China in 2022 and Master degree in Computer Science at National University of Singapore, Singapore, in 2024. He is currently working toward the Ph.D. degree in Computer Science with National University of Singapore, Singapore. His research interests include multi-modal understanding and computer vision.

Jinming Cao received her Ph.D. with the School of Computer Science and Technology from Shandong University in 2022. She is currently working as a Research Fellow with the National University of Singapore. Her research interests include computer vision and 3D vision.

Ting Wang received the BE degree in software engineering from Northwestern University in 2024. She is currently pursuing the MS degree in computer technology at Northwestern Polytechnical University. Her research interests include efficient fine-tuning and domain adaptation of large language models.

Yifang Yin is currently a Senior Scientist in the Machine Intellection Department, Institute for Infocomm Research (I^2R), A*STAR. She received the B.E. degree in the Department of Computer Science and Technology from Northeastern University, Shenyang, China, in 2011, and the Ph.D. degree in Computer Science from National University of Singapore, Singapore, in 2016. Her research interests include machine learning, multimodal multimedia analysis, spatiotemporal data mining, and AI for weather science.

Li Li is a PhD student at Thomas Lord Department of Computer Science, University of Southern California, supervised by Prof. Yue Zhao. His research interests are centered around multimodal learning, Out-of-Distribution detection, and information retrieval.

Shili Xiang is currently a Principal Scientist at the Institute for Infocomm Research (I^2R), Agency for Science, Technology and Research, Singapore (A*STAR). She received the BE degree from University of Science and Technology of China, Hefei, China, in 2003 and the PhD degree in Computer Science from National University of Singapore, Singapore, in 2011. Her research interests focus on spatiotemporal data intelligence and its diverse applications, including urban mobility and optimization, AI for transportation, and AI for weather science.

Ying Zhang received the B.E. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2009, and the Ph.D. degree in computer science from the National University of Singapore, in 2014. She is currently a professor with School of Computer Science of Northwestern Polytechnical University, Xi'an, China. Her research interests include Artificial Intelligence, Edge Intelligence and Crowd Collaboration.

Roger Zimmermann (Senior Member, IEEE) received the M.S. and Ph.D. degrees from the University of Southern California (USC), in 1994 and 1998, respectively. He is a Professor with the Department of Computer Science, National University of Singapore (NUS). He is also the Deputy Director with the Smart Systems Institute (SSI) and a Key Investigator with the Grab-NUS AI Lab. He has coauthored a book, seven patents, and more than 300 conference publications, journal articles, and book chapters. His research interests include streaming media architectures, multimedia networking, applications of machine/deep learning, and spatial data analytics. He is currently an Associate Editor for IEEE MultiMedia, Transactions on Multimedia Computing, Communications, and Applications (TOMM) (ACM), Multimedia Tools and Applications (MTAP) (Springer), and the IEEE Open Journal of the Communications Society (OJ-COMS). He is a distinguished member of the ACM and a senior member of the IEEE.